AFRL-HE-WP-TR-2006-0002

AIR FORCE RESEARCH LABORATORY

An Investigation of Human Performance Model Validation

Floyd Glenn
James Stokes
Kelly Neville
Kevin Bracken

CHI Systems, Inc.
1035 Virginia Drive, Suite 300
Fort Washington, PA 19034

March 2005

Final Report for December 2003 to March 2005

Human Effectiveness Directorate
Warfighter Interface Division
Cognitive Systems Branch
2698 G Street
Wright-Patterson AFB OH 45433-7604

NOTICES

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data, does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them

This report was cleared for public release by the Air Force Research Laboratory Wright Site Public Affairs Office (AFRL/WS) and is releasable to the National Technical Information Service (NTIS). It will be available to the general public, including foreign nationals.

Please do not request copies of this report from the Air Force Research Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161

Federal Government agencies and their contractors registered with the Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
8725 John J. Kingman Road, Suite 0944
Ft. Belvoir, VA 22060-6218

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-WP-TR-2006-0002

This technical report has been reviewed and is approved for publication.

FOR THE DIRECTOR

//SIGNED//

MARIS M. VIKMANIS
Chief, Warfighter Interface Division
Human Effectiveness Directorate

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
March 2005	Final	December 2003 - March 2005

4. TITLE AND SUBTITLE
An Investigation of Human Performance Model Validation

5a. CONTRACT NUMBER
FA8650-04-C-6438

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
63231F

6. AUTHOR(S)
Floyd Glenn, James Stokes, Kelly Neville, Kevin Bracken

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
28300410

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

CHI Systems, Inc.
1035 Virginia Drive, Suite 300
Fort Washington, PA 19034

8. PERFORMING ORGANIZATION REPORT NUMBER

04009.050329

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Cognitive Systems Branch
Wright-Patterson AFB OH 45433-7604

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

AFRL-HE-WP-TR-2006-0002

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited. PA cleared 2-21-06, AFRL/WS-06-0477.

13. SUPPLEMENTARY NOTES
AFRL/PA cleared 2-21-06, AFRL/WS-06-0477.

14. ABSTRACT
This report summarizes research to develop a methodology for calibration and validation of human performance models in support of simulation-based acquisition (SBA) processes - a human performance modeling validation program. A review was conducted of the various types of human performance models (HPMs) that have been developed and considered for SBA, focusing particularly on the category of cognitive models (CMs). The results of an informal review of the usage of HPMs and CMs in several major recent defense acquisition programs are described. Also, details the conceptual framework based on an investigation of the characteristics of a wide variety of performance modeling frameworks and application domains. Offered is the initial taxonomies of model actions and empirical performance actions that will support the necessary mappings between model predictions and empirical observations for all models over the full range of required representation detail, and across all stages of the acquisition process in order to establish a solid analytic basis for calibration and validation of SBA processes and human performance modeling frameworks. Also, described are the methods for specifying performance measures so that for any given design decision, performance measures captured using a model can be mapped to performance measures obtained during live test and evaluation.

15. SUBJECT TERMS Human Performance Model, Cognitive Model, Simulation, Validation, Calibration, Simulation-based Acquisition, Virtual Environment, Human-system Integration

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Lt Matthew T. Eaton
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR	76	19b. TELEPHONE NUMBER *(include area code)* 937-656-7001

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. 239.18

THIS PAGE LEFT INTENTIONALLY BLANK

TABLE OF CONTENTS

TABLE OF CONTENTS III
LIST OF FIGURES IV
LIST OF TABLES IV
ACKNOWLEDGEMENTS V
1. INTRODUCTION 1
2. SBA: THE BACKGROUND 4
2.1 Human Performance Models (HPMs) 4
2.1 SBA 6
2.2 Service Directives on SBA 7
2.3 Prior DoD Programs to Employ HPMs in SBA 8
3. HUMAN PERFORMANCE MODELING AND COGNITIVE MODELING IN CURRENT ACQUISITION PROGRAMS 11
3.1 Joint Strike Fighter (JSF) Program 13
3.2 Joint Synthetic Battlespace (JSB) Program 15
3.2.1 Project Gnosis 17
3.2.2 Cultural Modeling for Command and Control (CMC2) 17
3.2.3 The Role Player Intelligent Controller Node (RPICN) 17
3.2.4 Joint STARS Operator Surrogate Human (JOSH) 17
3.2.5 Obstacles to CM and HPM Use in SBA 18
3.2.6 Advances Impacting Future CM and HPM Use in SBA 20
3.3 Joint Unmanned Combat Air Systems (J-UCAS) 21
3.4 New Navy Ship Class -- DD(X) 21
3.5 Army Future Combat System (FCS) 22
3.6 Army SMART 23
4. HPMS, CMS AND ALTERNATIVES IN THE CURRENT ACQUISTION PROCESS 27
5. VALIDATION OF SIMULATIONS AND COGNITIVE MODELS 31
5.1 Application Validity and SBA 31
5.1.1 Construct validity 31
5.1.2 Qualitative assessment of construct validity 32
5.1.3 Quantitative assessment of construct validity 32
5.1.4 Application validity 33
5.2 General issues in CM validation 35
5.3 Concepts of Calibration vs Validation 38
5.4 Aspects of HPM and CM Validity 41
6. CM VALIDATION IN THE EMERGING SBA CONTEXT 50
6.1 Anticipated Relationship of CM to Virtual Environment 50
6.2 CMs and Application Validity 52
6.3 Validity as Correspondence 53
6.4 Calibration and Validation of CMs 54
6.5 Developing and Validating CMs in the VE Context 56
7. CONCLUSIONS 60
8. REFERENCES 61
9. ACRONYM LIST 67

List of Figures

FIGURE 1. OVERVIEW CONCEPT FOR HPM VALIDATION METHODOLOGY 3
FIGURE 2. CONCEPTUAL SPACE FOR HUMAN PERFORMANCE REPRESENTATION 5
FIGURE 3. ENVISIONED DEVELOPMENT AND USE OF HPM MODELS IN SBA 7
FIGURE 4. LAYERS OF MODEL GENERATION & SPECIFICATION 46
FIGURE 5. COGNITIVE CAPABILITIES SPECIFIED BY THE HUMAN BEHAVIOR MODEL 46
FIGURE 6. ARCHITECTURAL VIEW OF HOST AND CM SIMULATIONS 51
FIGURE 7. VALIDATION DATA POINTS IN THE DESIGN SPACE 56
FIGURE 8. CM VALIDATION IN THE SBA PROCESS 59

List of Tables

TABLE 1. SCALE FOR ASSESSING THE VALIDITY OF STUDIES CONDUCTED TO DERIVE HPMS AND CMS FROM SILVERMAN ET AL., (REFERENCE 70) 45
TABLE 2. CONTEXT FACTORS THAT MAY IMPACT PREVIOUSLY ESTABLISHED VALIDTY OF A CM OR HPM 48

Acknowledgements

The reported work was performed under USAF contract number FA8650-04-C-6438 with the Air Force Research Laboratory (AFRL), Human Effectiveness Directorate (AFRL/HE). The authors wish to thank 1st Lt Matthew Eaton, who has provided valuable advice and assistance in this project. The authors would also like to acknowledge the important guidance and technical contribution of Dr. Michael Young (also of AFRL). All judgments and opinions are, however, those of the authors and do not represent positions of the U.S. Air Force.

THIS PAGE LEFT INTENTIONALLY BLANK

1. Introduction

The technology of building models and simulations of human performance has made rapid advances over the past few decades to the point that there is now considerable interest in using such models to evaluate human-system integration (HSI) aspects of new advanced technology systems. This is particularly the case for complex, large-scale, weapons systems where it would be appealing to be able to use human models in place of real human operators or maintainers in order to determine if the intended users will be able to use/operate/maintain the system adequately. However, consideration of using any human performance model (HPM) for design evaluation quickly raises the question of the validity of the HPM. There seems to be a general sense, at the same time, that there is not any very good evidence concerning the validity of any HPMs, or at least that there is no professional consensus as to what procedures should be employed to assess HPM validity for most types of HPMs. This report investigates the problem of validation of HPMs, seeking to understand why validation of models in this area should be so different than the situation with other engineering models, and also seeking to recommend appropriate methods for systems acquisition programs to employ in using and concurrently validating HPMs. We will refer to this program occasionally as the MODVAL program.

The initial goal of this program was to develop a methodology for validating HPMs used in simulation-based acquisition (SBA) applications using the data that is typically generated in test and evaluation activities. However, since it was determined midway through the project that there were virtually no identifiable cases of HPM uses of the primary types of interest for SBA applications, this goal was adjusted. The revised goal of this project is to provide guidance for the use of HPMs for simulation-based acquisition (SBA) applications.The purposes of the HPMs are to support and enhance system performance predictions/evaluations that depend in some way on human behavior or performance and that must be conducted in the course of deciding between system design alternatives in a succession of levels of detail. We need to know that each model-based prediction/evaluation is good enough so that we make the right design decision with appropriate confidence. But we also want to know that we are not expending any more effort (and other costs) than necessary to ensure that the right decision is supported at each stage. Ultimately, we want to enable SBA managers to have confidence that they are selecting the right HPM tools and using them correctly in order to make efficient and optimally informed SBA decisions.

Thus, the central objective of the present methodology development effort has been to support validations and calibrations of models of complex task performance for SBA so that the results of each validation effort can be used incrementally to inform subsequent decisions about what modeling tools and techniques are most appropriate for each new SBA activity. We have investigated how to attribute validation results to all of the distinct facets of SBA modeling efforts so that for subsequent SBA problems we can better determine what technique should be used, what kinds of analyst skills are needed, what component model elements should be incorporated, and how the free parameters of the models should be estimated. Clearly, this also entails the development of new guidelines for the collection and analysis of performance data from both models and empirical activities, and the construction of scenarios that will sufficiently exercise model and human participants.

Development of this methodology necessarily begins with an investigation of SBA requirements that are intended to be addressed by HPM tools. Following that, we survey and analyze the various HPM tools and techniques that are available to support these types of SBA needs. Next, we review techniques for comparison of simulation-generated data with empirical T&E data in order to adjust and calibrate the simulation models and to develop conclusions and diagnostic inferences regarding validation. Then we investigate the kinds of data collection that are likely to be feasible in conjunction with military system T&E activities.

Throughout this review, we are repeatedly confronted with the necessity of considering the various alternatives for using HPMs to support SBA decisions. A principal alternative of interest is the use of experimentation in which real candidate users/operators interact with a prototype version of the system design in some moderately realistic mock-up or simulation environment. The use of virtual environment (VE) simulations to conduct human-in-the-loop (HITL) experiments is generally seen as the most cost-effective option for such experiments for most major DoD systems. It is also noteworthy that these same types of VE-HITL experiments also constitute the primary potential source of empirical data for HPM calibration and validation. Thus, it is clear that the VE-HITL options are necessarily closely related in complex ways with the HPM usage and validation issues. Accordingly, we offer some observations about the appropriate and expected coordination of the use of these technology options in support of SBA. The overall concept of the intended methodology is illustrated in flow-chart form in Figure 1 as integrated into a skeletal system acquisition process with SBA support. The boxes in the figure that comprise the primary focus of the current efforts are shaded in light blue and are further discussed in subsequent sections of this report.

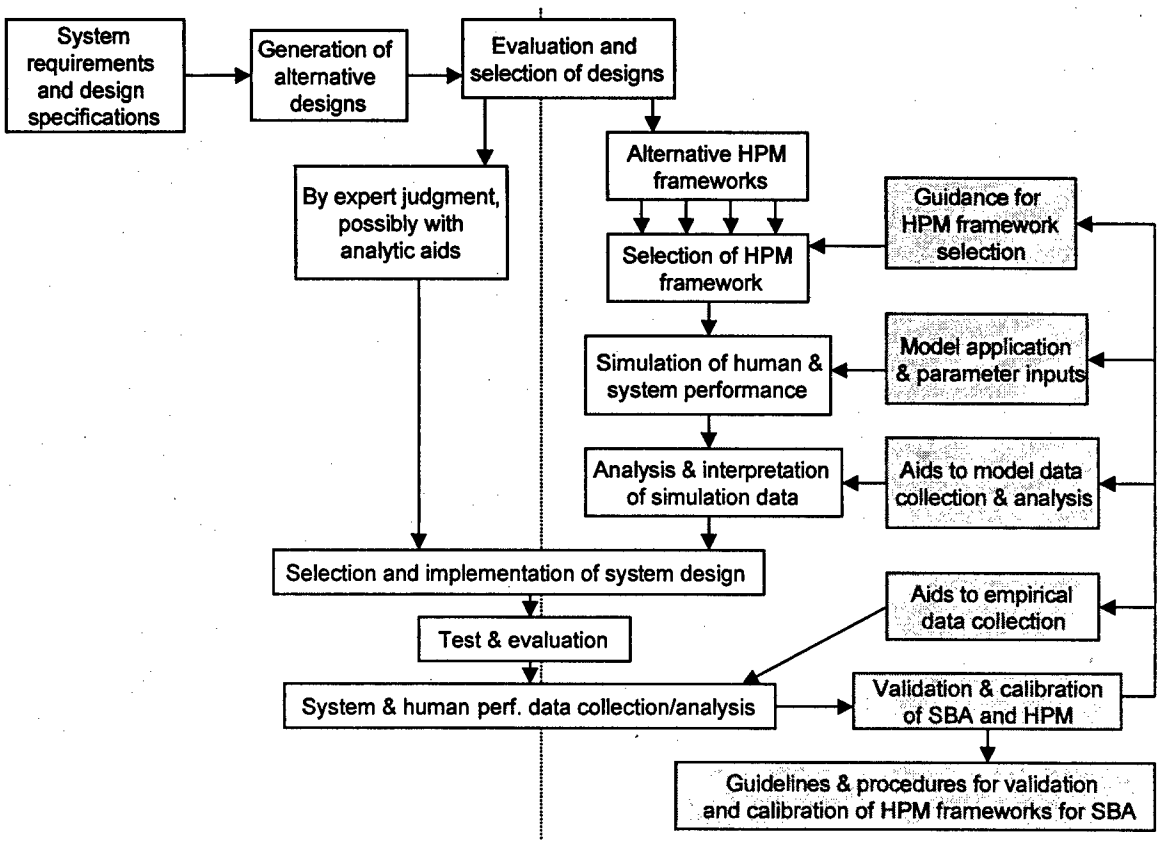

Figure 1. Overview Concept for HPM Validation Methodology

2. SBA: The Background

The various technologies associated with simulation of system dynamics have been developed for three principal purposes – to aid in the design of systems, to aid in the training of users of the systems, and to support analyses of system performance for other purposes (such as tactics development). We are presently concerned mainly with the design applications where simulations are used for analytic, diagnostic, and evaluative purposes, enabling the designer to develop, examine, and evaluate designs without having to commit to the costs and delays of actual implementation of the candidate designs. Simulations can serve to represent many different perspectives and levels of scope for a system concept according to the character of the system and the objectives of the simulation developer. Representation of the human performance of system users (or operators, participants, human elements, etc.) is only one aspect of system simulation, and indeed human simulation is not necessarily an explicit focus of most system simulation work. Still, human performance simulation must be understood within the broader context of system simulation in general, and it is this broad context that is addressed by the concept of simulation-based acquisition (SBA). Accordingly, we will review the various types of representations of human performance available for SBA, as well as the concept and current status of SBA in the U.S. Department of Defense (DoD). The section which follows this general background on SBA, will provide brief background reviews of a few of the DoD efforts that have been pursued to date to integrate HPMs into SBA processes since it is through these efforts that the current interest in validation of HPMs has developed.

2.1 Human Performance Models (HPMs)

The range of existing and possible HPMs is broad and complex corresponding to the many components, layers, and perspectives on human performance. A notional representation of this space of human performance representation (HPR) is provided in Figure 2. The disc in the figure represents the primary functional areas for the HPR, with more basic functions near the center and more highly integrated functions toward the periphery. Differences in representation of the basic functions of perception, physical form & action, declarative knowledge, and procedural knowledge serve to characterize some of the major differences between the various types of current HPMs. The four vertical arrows through the disc represent additional HPR dimensions that seem to be orthogonal to the basic functions in the disc. Attention/automaticity and affect/moderators both serve to represent HPR aspects with diffuse connections to the basic resources in the disc. The dimensions of competence vs. performance and general vs. domain-specific address two different aspects of behavior organization.

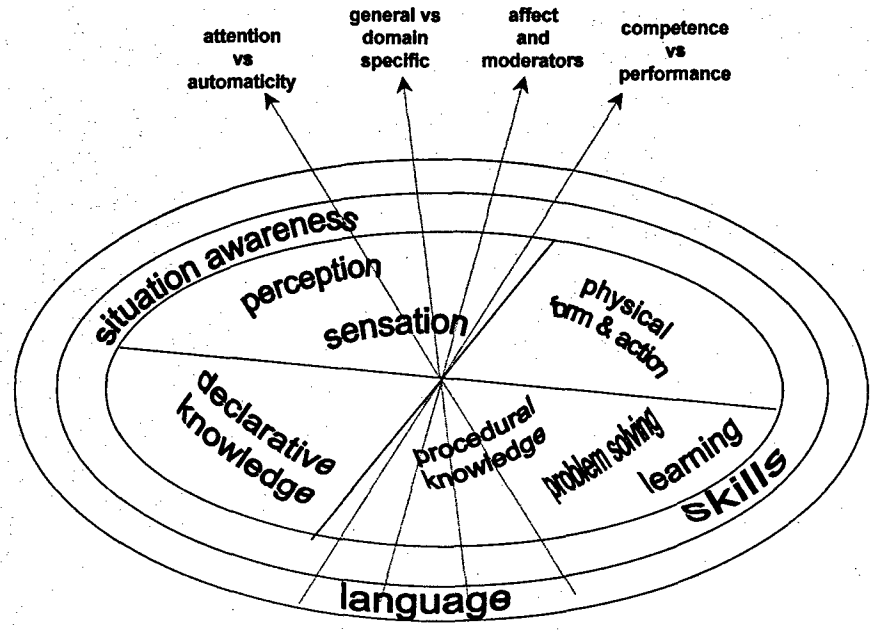

Figure 2. Conceptual space for human performance representation

Although it is our view that each existing HPM occupies a distinct and different place in the space of Figure 2, it is useful to distinguish some of the major categories of HPM:

- Biodynamic models (e.g., Jack, Ergo, and Combiman) are defined almost completely within the "physical form & action" segment.
- Control theory models (classical and optimal) combine aspects of perception and physical action.
- Task network models represent a subset of procedural knowledge.
- Cognitive model (CM) architectures incorporate representations for each of the basic four segments of the disc and occasionally extend into the peripheral areas of skills and SA and more rarely into language.

And, of course, there are many other types of HPMs that address just one segment of this space, especially in the areas of perception and language. With regard to CMs, it is useful to recognize that each of the several available CM architectures and frameworks address somewhat different portions of this HPR space, offering more or less detailed representations of perception, action, problem solving, and learning processes. Still, all CM architectures provide some level of representation in the four basic functional areas of perception, action, declarative knowledge, and procedural knowledge.

Recent reviews (References 1, 2) have identified and compared some of the principal exemplars of the CM category, which includes ACT-R, COGNET/iGEN®, EPIC, OMAR, and SOAR, among others. It is noteworthy that cognitive models that are constructed within any of these frameworks for DoD program applications are typically not implemented in isolation but rather within the context of some host simulation for the environment of interest, such as an aircraft or ship in a warfare environment. It is also worth recognizing that a CM could also be constructed without the use of any of these general-purpose CM architecture environments, but rather in a general-purpose programming language (such as LISP), but we could also consider

most such languages as sorts of default or degenerate CM architectures; in the remainder we will assume that some explicit CM architecture is employed in all cases of interest.

2.1 SBA

The DoD system acquisition process is based on a standardized sequential process that is designed to insure the correctness and quality of all aspects of the process. As DoD systems have grown increasingly large and complex over recent decades, however, so has the standard acquisition process, to the point where it encompasses mind-boggling volumes of directives, standards, specifications, and other process documentation, and the complexity of the process results in system development timelines than can stretch over more than twenty years from the specification of the original concept requirement to the beginning of full-scale production of operational systems. There have naturally been many refinements and revisions to the standard acquisition process including expediting abridgements, abandonment of government standards in favor of commercial standards, incorporation of commercial off-the-shelf (COTS) products, and replacement of strict sequential development with an iterative spiral development process. Still, despite all of these refinements, the major stages of development are inescapable in most cases, starting with a determination of a new required operational capability (such as by the DoD Joint Requirements Oversight Council, JROC, working with the Joint Capabilities Integration and Development Systems, JCIDS) they include:

- Concept formulation and refinement
- Technology development
- System development and demonstration
- Production and deployment
- Operations and support

Although earlier applications of simulation technology were envisioned to support just the first two or three of these acquisition stages, the more recent concept of spiral development for acquisition has made it clear that concept refinement and technology development can still be revisited in the later stages, so simulation can be useful throughout all stages.

A common distinction that is made between usages of simulation is that between constructive and virtual applications. A constructive simulation is one in which all relevant elements, including any human participants, are represented by simulation, whereas a virtual simulation is one in which the simulation serves to provide the representation of the non-human elements of the system and environment along with a simulated interface for the human user, thus allowing the human to interact with the simulated system and environment as if it were real. In actuality, there are many gradations and combinations of constructive and virtual characteristics in cases where multiple human participants can be simulated in various levels of detail or represented by the control of human participants who could operate through virtual interfaces with varying degrees of fidelity. Figure 3 illustrates the nominal sequence of acquisition development stages (top third of the figure), the expected mix of constructive simulation, virtual simulation, and live prototype testing that are expected for T&E purposes (middle third of the figure), and the corresponding evolution of the specificity of system design concepts, HPMs, and T&E applications.

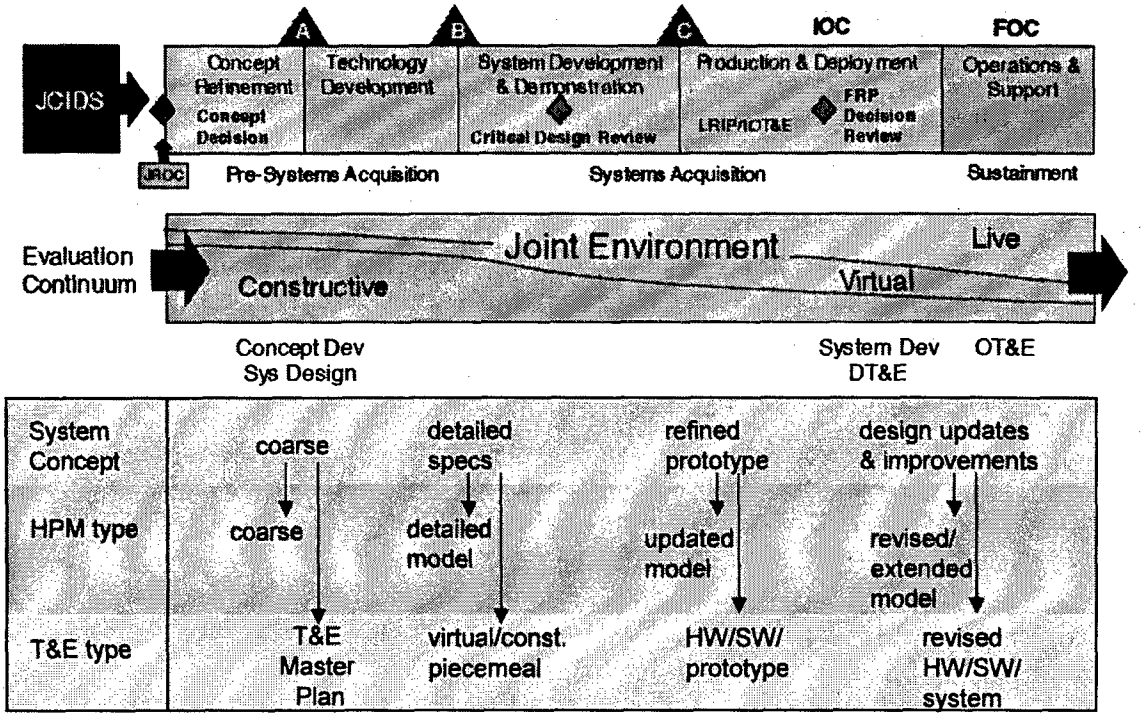

Figure 3. Envisioned development and use of HPM models in SBA (top two-thirds of figure taken from Reference 3)

2.2 Service Directives on SBA

The highest level specification of DoD policy for defense system acquisition is specified in DoD directives 5000.1 and 5000.2 (i.e., DOD 5000.1, 5000.2). The latest revisions to these directives make clear the requirement to develop and employ a robust and effective SBA approach to system acquisition, though different approaches are being pursued by each of the services (Reference 4). The role of Modeling and Simulation (M&S) in support of this approach is further specified in DOD 5000.59 and 5000.61. This DoD policy further requires each service component to develop M&S policies and procedures that are consistent with their service-dependent needs. For example, the Air Force Policy Document AFPD-16-10 tailors the DoD M&S instructions to its needs. More directly relevant to the objectives of our effort, the Air Force further defines Validation, Verification and Accreditation (VV&A) policy and procedures in AFI-16-1001.

In support of and consistent with these general policies, we are naturally interested in providing a methodology and tools to support a broad range of likely upcoming SBA efforts in association with major new military system acquisitions, not just one or two immediate cases. The diversity across different military systems is considerable in many regards, including especially the kinds of human performance issues that are relevant for critical SBA decisions. In some cases, such as for fighter cockpits, quick visual and manual access to displays and controls can be very important and difficult to achieve, whereas other military workstations pose few concerns in that area. Similarly, some systems present highly stressed workload conditions,

whereas others do not. It is important for us to identify a comprehensive range of the kinds of human performance issues that are of concern for SBA decisions across the full variety of military systems. Based on the enumeration of the kinds of military systems to address, we must identify what aspects of performance we want to assess (e.g., performance times, decision quality, workload, situation awareness, confidence, etc.), what performance-influencing factors we want to be able to address (e.g., stress, fatigue, vibration, noise, etc.), and what kinds of individual differences in performance characteristics we need to be able to represent for what populations (e.g., expertise, intelligence, personality, culture, etc.). We must also address the accuracy of prediction that is desired to support the SBA decision, for example by determining how sensitive the SBA decisions are likely to be to variations in predicted performance times or workload scores.

2.3 Prior DoD Programs to Employ HPMs in SBA

The idea of using HPMs in SBA has actually been with us for a long time, at least since the early 1960s when the Siegel-Wolf task network models were applied to evaluate designs of several major military systems (Reference 5). Whereas the Siegel-Wolf models were developed by employing a general conceptual methodology in order to implement each model in its own unique code for simulation software (Fortran at the time), subsequent programs in the Navy (with the Human Operator Simulator, HOS; Reference 6) and in the Air Force (with the Systems Analysis of Integrated Networks of Tasks, SAINT; Reference 7) sought to develop general-purpose software tools and environments in order to simplify, aid, and standardize the human performance modeling activities for support of SBA. However, the SBA experience in use of HOS and SAINT, as with the earlier Siegel-Wolf models, was that the costs of using these models was high (with one or more person-year of effort required for typical model applications) and the predictive accuracies were indeterminate and possibly erratic. At least part of the problem was that various other more basic HPMs were typically embedded within the broadly scoped applications of tools like HOS and SAINT, each of which introduces additional potential for inaccuracy and unreliability.

There has also long been recognition in SBA communities that it might be appropriate to establish collections or families of complementary HPMs to be used for different aspects and issues of SBA. Such collections were developed by the Navy in its CAFES program and by the Air Force with its CADET program, both in the 1970s, with some components addressing task performance (like HOS and SAINT), others anthropometry (like the Air Force COMBIMAN and the Navy CAR), etc. In the militarily important task domain of manual tracking of dynamic targets (fundamental for many aspects of aircraft piloting and air defense weapons operations), a fairly elaborate genre of mathematical models has been developed in order to predict human performance characteristics from detailed design features of control systems along with environmental dynamics (e.g., Reference 8). There has also been a long tradition in the development of models for factors that seem to have diffuse moderating effects on virtually all human performance, factors such as fatigue and circadian rhythms, stress, chemical and biological agent effects, performance enhancing/degrading drugs, etc. (e.g., Reference 9).

Over the past decade, the Army has sought to impose some order on the otherwise seemingly arbitrary processes of selecting and applying human performance modeling tools in

support of SBA. In their IMPRINT program (Reference 10), they have commissioned the development of several simulation-based tools to address different distinct aspects of human accommodation and performance issues pertinent to SBA decisions, including components to address issues associated with personnel selection, training, survivability, workload, and system manning. Use of these tools has been required in the course of recent Army weapons system acquisition contracts. The Air Force has recently implemented its own variant of the IMPRINT task performance modeling tool for its CART program (Reference 11). This tool employs a task network representation, directly descended from the SAINT tool, to generate dynamically executable descriptions of human performance. This task modeling tool, like all other task network simulations of human performance back to Siegel-Wolf, also offers a library of micro-model and moderator functions to support the construction of human performance simulations. Without any micro-models, the analyst building the simulation would have to specify all of the relevant performance characteristics of each task (e.g., time duration, prerequisites, post-task branching, workload characteristics, etc.). Micro-models provide parametric descriptions of some or all of these parameters. For example, a target detection micro-model might establish a deterministic or probabilistic prediction of the time required for target detection according to scene characteristics and human operator state characteristics. Moderator functions, such as for fatigue or stress, might establish proportional adjustments to all task times, or may possibly make differential adjustments to different tasks or different micro-models within tasks. Another noteworthy characteristic of the IMPRINT task modeling tool is that it offers a substantial library of generic simulation templates for a broad range of typical military weapons systems, such as for a fighter aircraft pilot, tank commander, or anti-aircraft gunner, thus facilitating construction of new applications by starting from a generic template that is fairly close to the concept for the SBA system of interest.

In its recent Agent-based Modeling and Behavioral Representation (AMBR) program (Reference 12), the Air Force has also investigated the relative effectiveness of several competing cognitive modeling architectures for the simulation of detailed human performance characteristics in the context of an abstracted air traffic control task that is similar to many military task environments. As distinct from the task network representations of IMPRINT and CART, AMBR has investigated only a small set of alternative knowledge-based models of cognitive performance involving complex architectures that serve to integrate component functions of attention, memory, perception, decision-making, motor action, and so on. AMBR has investigated the relative effectiveness of four different cognitive modeling architectures (ACT-R, DCOG, EASE, an ACT-R/Soar/EPIC hybrid, and iGEN®) as used by the teams who developed each architecture for predicting a broad range of performance characteristics in the chosen task environment, including the fine details of task dynamics, learning of complex concepts, and transfer of learning to new conditions.

The issue of model validation has been addressed throughout all of these HPM developments from the very beginning. But curiously, there is little that can be said definitively about the validity of any of the complex task performance models (i.e., the task network models and the cognitive architectures), even within the constraint of the specific SBA efforts in which they have been seriously applied. As noted by Young (Reference 13), different kinds of HPMs present different issues and options for validation. And in addition to the various types of HPM, it is also appropriate to recognize that validation efforts must probably be tailored to a defined

range of application environments and usages (with major differences across SBA, training system, and decision support system applications) as discussed recently by Campbell and Bolton (in press) under the concept of "application validity." In some cases, human performance data has been collected on a working implementation of the system being modeled and then compared to model data, resulting in judgments being made regarding the closeness of the correspondence. For simple task models such as manual control models (e.g., Reference 8) and for anthropometric models (e.g., Reference 14), these types of validations have been extremely productive and conclusive. The problem for models of complex task performance is that we do not have clear criteria for acceptable prediction accuracy for any of the many aspects of performance that these models might predict (e.g., performance timelines, decisions at choice-points, corollary behaviors such as eye movements, affective characteristics such as workload and stress, etc.). Furthermore, because of the complex character of these models, it is generally impossible to attribute any part of the correspondence between model and empirical data to any particular component, layer, or other aspect of the model application. Thus, it is difficult to infer how to attribute the results of any validation effort to the suitability of the task network or cognitive modeling architecture, or to the skills of the modeling team that constructs each application, or the particular collection of micro-models and moderator function models that are employed for that application, or to the technique employed for the estimation of parameter values for all of the many free parameters incorporated in the model application. Also notably absent from most validation efforts is any systematic consideration of the realistic available alternatives for making the SBA decisions of concern, occasionally considering alternative models as in the AMBR program, but seldom identifying and evaluating the non-model-based analytic techniques that the system analyst might use to inform the same design evaluation decisions.

3. Human Performance Modeling and Cognitive Modeling in Current Acquisition Programs

With regard to the issue of active use of CMs in SBA, there are a number of reasons to expect that some DoD acquisitions would be using CMs in a systematic fashion. And, in fact, there may be some such CM applications in some DoD programs, but these cases have been difficult to find. Reasons to expect that CMs might currently be in active use in this way include:

- Publications of research and plans for such applications have been widely disseminated in the DoD human factors community for over twenty years (References 15, 16).
- The Army has mandated the use of standard workload modeling tools for Army system acquisitions (References 10, 17).
- The Navy has conducted extensive research efforts regarding SBA use of human performance models in preparations for a new fighter aircraft (through the Advanced Technology Crew Station program; Reference 18) and for the DD-X ships (through the Manning Affordability Initiative; Reference 19).
- The noteworthy success of NYNEX in their application of a GOMS model for evaluating a new telephone operator system design in Project Ernestine (Reference 20) would seem to be a solid indication that CM technology is sufficiently mature for SBA use.

On the other hand, the only mandated modeling that we could identify is workload modeling in support of Army MANPRINT aspects of Army acquisitions, and workload modeling does not require use of a CM. The type of workload modeling that is specifically recommended for Army MANPRINT applications is based on the use of a task-network modeling framework combined with subjective estimates of workload for each of the tasks (or subtasks). While it is possible to use a CM to generate workload estimates and avoid much of the requirement for subjective estimates (e.g., Reference 21), we have found no evidence that anything other than task networks and subjective estimates have been used in the Army MANPRINT cases.

Two compelling reasons to believe that CMs may not have yet been used in any major way on any acquisition program are that:

- Any SBA application of a CM would require that the acquisition program would have to develop and maintain a simulation of all aspects of the system and environment with which the human operator must interact, and the CM would have to be integrated with such a host simulation – but simulations with all of these characteristics are not readily available for most acquisition programs, though, as we will discuss later, they are becoming increasingly common for some system components.
- The acknowledged high cost of CM development would make it unlikely that an effort of such magnitude would go unnoticed.

If there is ongoing use of a CM in the course of a major DoD acquisition, it is also possible that proprietary protection concerns have prevented us from learning about it. If the acquisition is still in a competitive stage, then the prime contractor may not want its methods to be divulged to any competitors. Or the prime contractor might not be willing to commit publicly to the use of the CM when the results or quality of the results are still in doubt.

In any case, it seems evident that this project could not count on finding any systematic SBA application of a CM to use as a case study. But this has not represented a serious impasse for the effort. Rather, it has caused us instead to select and examine acquisition case studies in order to identify points in the acquisition process where CMs could be reasonably considered as candidates for making acquisition decisions and what the criteria should be for deciding whether or not to use a CM or any other tool or information source for that purpose. In fact, we had some discussions in this vein with the JSF engineers when we were searching for evidence of CM use on that program. But even if we found that one or more CMs were being used in some ongoing acquisitions, those cases would not be likely to be addressing all of the ways that the CM might be used to support the acquisition, nor would they necessarily provide a clear record of how the decision was made to use the CM. Accordingly, we turned from a search for active SBA applications of CMs to a review of any SBA acquisitions in order to identify all relevant opportunities and criteria for use of CMs in support of the acquisition process.

With regard to identification of acquisition programs using CMs, we recognized from the outset that it would not be possible to conduct a systematic review of all, or even many acquisition programs. Rather, we expected to find a few relevant programs, actively using CMs that we could investigate in some detail. JSF was the first, and it seemed to proceed according to plan through the early contacts and discussions. However, it became clear in a telecon with the JSF team that they were not using any CMs, or any other HPMs for that matter, for any HSI applications for the JSF aircrew (though they did report some use of biodynamic models such as Ergo and Jack for maintainer applications). This telecon followed earlier rumors and indications from the same JSF people that they were using HPMs that they would be willing to discuss with us. It was only when the discussion evolved to the point of exploring detailed model uses that we discovered the disconnect. Not only were they not using any CMs for design evaluations, they were not even using any CMs for simulation of other peripheral role players for virtual environment simulation for empirical evaluation of design concepts. Rather, they used SMEs for both the primary and peripheral players in the virtual environment evaluations.

There have been various informal reports that the Navy's new DD(X) ship program was making extensive use of HPMs and CMs. However, in an informal discussion with one of the major subcontractors regarding their use of HPMs for design evaluation of a major ship component where the contractor had reported to the government that they were using biodynamic models, we learned that they had bought all of the major relevant COTS products (Ergo, Safework, and Jack) but had no intentions of actually using any of them. Unfortunately, our attempts to obtain further information from the current DD(X) program have been frustrated by the proprietary concerns of the contractors and the difficulty of obtaining the cooperation of the Navy program office.

Coincidentally, we have also had the opportunity recently to provide HSI subcontract support to one of the teams that is developing the Navy's new Littoral Combat Ship (LCS). LCS is very different from DD(X) in that it is a rapid, low-cost, legacy-focused acquisition, making maximum use of COTS/GOTS components. Nevertheless, it was very interesting to find that this program still produced an HSI plan that identifies several HPMs that it intends to employ, with descriptions that make it sound like these might involve true cognitive models. In fact, we

have learned from our direct involvement in the program, that the only modeling being done is some very crude task-network workload modeling and some biodynamic modeling of maintainers.

We have drawn some conclusions from these experiences that have guided our investigation:

(i) It is very difficult to determine how extensively and how seriously HPMs and CMs are being used in an acquisition program. It is hard to obtain access to the right people and obtain the kind of cooperation that is needed to dig deep enough into the program to identify and separate the claims from the realities. It is also possible that, when we have found no current uses of HPMs in the program, there may still have been some such applications in the earlier stages of acquisition when none of the current players (contractor and government) were involved.

(ii) We suspect that there may have been no serious applications of CMs for SBA design decision-making on any acquisition programs to date. Most of the large, high-profile programs will probably claim to use M&S for HSI. And most of them will probably acquire some tools and capabilities for that purpose. But we expect that most, if not all, are not currently following through on actual use of these models, especially not in the CM category, to make any real design decisions. (However, there is evidence of active use of CMs such as TacAir-Soar to represent peripheral players/platforms in simulation-based exercises that are used to evaluate system design concepts and tactics.) We expect that the acquisition programs (really the contractors) are obtaining and trying out the tools (presumably with the best of intentions), but then determining that they cannot justify any serious applications for design evaluation. We suspect that if there were any genuine use of a CM for any kind of real design evaluation, then everyone involved would make sure that the effort was highly publicized. The NYNEX Project Ernestine (Reference 20) effort is the primary data point in this arena and, despite the commercial sensitivity of the project, it was widely published.

(iii) The main obstacle to use of HPMs for SBA may, in fact, be the formulation of the kind of planning guidance discussed in Section 6 below.

The sections which follow present a review of major acquisition programs investigated as part of our research, noting their employment of HPMs and CMs, and any efforts toward HPM and CM validation.

3.1 Joint Strike Fighter (JSF) Program

The Joint Strike Fighter (JSF) program (see References 22, 23) originated in the early 1990s as a result of the reorganization of several existing DoD tactical aircraft and technology initiatives. In 1994, following numerous trade studies, Boeing, Lockheed Martin, McDonnell Douglas, and Northrop Grumman were each awarded fifteen-month Concept Definition and Design Research (CDDR) contracts. In 1996, JSF was designated an Acquisition Category I, DoD acquisition program with contract awards to two consortia, led by Boeing Aerospace and

Lockheed Martin, for the Concept Demonstration Phase of the effort. In 2001, a team led by Lockheed Martin was awarded the contract to build JSF. The Critical Design Review is now scheduled for April 2005, with the first F-35A having begun airframe assembly and scheduled for its initial flight in August 2006.

In addition to reviewing available JSF documentation, discussions were held with representatives of the JSF prime contractor and the JSF program office in the course of the MODVAL research. The possibility that the JSF program is employing human performance modeling technologies was explored in several areas, including issues of accommodation for pilots, maintainers, and production workers.

In general, there are three types of pilot accommodation issues to be addressed: cognitive, perceptual, and physical. JSF mainly uses models to address issues of physical accommodation, with some attention to perceptual accommodation. Cognitive accommodation is being addressed during rapid prototype evaluations. For example, issues of Situation Assessment/Awareness are addressed during human-in-the-loop prototype and full-emersion simulation evaluations. Pilot workload is the focus of a separate team at BAE Systems in Fort Worth, a part of the larger Lockheed team. The team plans to revitalize its workload working group, possibly looking at cognitive accommodation, SA, etc. Based on the discussions held with JSF engineers, no human performance models are currently being used to address issues of pilot accommodation.

The engineering group working on weapons loading, for example, has been using the DELMIA SAFEWORK and Ergo products (Deneb & SAFEWORK are now part of DELMIA). ENVISION digital manufacturing software (see References 24, 25), also from DELMIA, is being used as an ergonomics visualization platform, allowing engineers to view the ERGOman or ERGOwoman mannequins as they perform production tasks in a constructive simulation, in support of F-35 assembly. Software developed at Sandia National Laboratories is used to analyze movements and create the most efficient path for a given task.

The Ergo and SAFEWORK products are seen as tools to address anthropometry issues, not as models. The models here are the "case descriptions" (tall/heavy, short/light, etc.). Lockheed's customer has asked about the validity of Ergo and the other tools. The developers' answer to this request was given in terms of team confidence in the tools, based on years of use, although validation is currently being pursued.

In all areas of the design/development process, the JSF project quickly moves to human-in-the-loop studies, based on physical mockups. When adversaries are involved in the studies, for example, they are not provided as human performance models. They are included by means of lower-fidelity enemy platforms manned by trained human role players. (The cockpit development team includes many pilots who would be averse to using models in place of human-in-the-loop testing.) Even the weapons loading group has already followed up their modeling work with real maintenance work.

Under the reformed acquisition process the government no longer tells the developer "how to do it" or even "what to do." There is no Human Engineering Design Approach Document – Operator (HEDAD-O) requirement, for example (whereas a HEDAD-O was

required for acquisitions of all human-operated systems prior to the reform). Nor is any kind of SBA requirement imposed on the developer. Frequently working under severe time pressures, the JSF developers clearly perceive human-in-the-loop studies as their most efficient option, limiting human performance modeling to the arena of physical accommodation issues in the initial stages of rapid prototyping.

3.2 Joint Synthetic Battlespace (JSB) Program

The *Joint Synthetic Battlespace* (JSB) is a high-fidelity simulation environment developed to support warfighter planning and training and, more pertinent to the present MODVAL effort, to support the development of new system designs, hardware, algorithms, pilot–vehicle interfaces, and tactics. Thus, one of the primary objectives of the JSB is to support two communities in parallel—the operational community and the acquisition community— and to do so using the same simulation environment. In support of acquisition, the JSB is intended to contribute to all acquisition stages, including concept definition, development, test, deployment, and sustainment. By providing a means for concepts and evolving technologies to be tried out in a realistic version of the environment in which they will be employed, final products should be relatively free of surprise glitches. As examples, access to this environment should prevent the production of technologies that cannot communicate with other key systems and the selection of information flow designs that do not support the work requirements of the operators.

A second objective of the JSB is to enhance simulation-based training and acquisition by providing unprecedented realism. This realism is to be achieved by means of an extensive collection of highly accurate system simulations complemented by validated and accredited models of the many complex elements that shape the warfighting environment, such as weather effects, terrain effects, weapons capabilities and effects, sensor performance, processing latencies, communication latencies, and so on. The simulations and models are integrated into the JSB architecture, which additionally is able to integrate with real systems.

There are at least three reasons why CMs may be considered an important component of the JSB—that is, three reasons why, in this particular acquisition environment, HITL simulation will not suffice and CMs, despite their typically higher initial cost, seem to be required. These reasons are discussed below and have to do with the JSB objective of achieving a high degree of realism, the complexity of the simulated joint battlespace, and the types of concepts and systems that are expected to be developed using the battlespace.

First, because a battlespace implies a large network encompassing many organizations and warfighters, the number of human roles to be represented is necessarily large and calls for a computer-based form of representation. The cost associated with using humans to fill these roles is most likely much higher than the cost associated with building CMs, each of which can be replicated to provide any number of synthetic role players of a given type for the JSB.

Second, the human operators who populate the joint battlespace environment are impacting it as much as modeled environmental and system factors such as the terrain, weather, and communications network characteristics, if not more so. This suggests that to achieve the highly accurate simulation environment that is sought, CMs are needed to produce realistic operator

performance, including realistic operator decision latencies, decision variability, workload effects, information overload effects, and event detection latencies and accuracy. CMs similarly are needed to produce realistic representations of enemy operator performance.

Third, the use of the JSB to evaluate C2 technologies, concepts, and processes has been emphasized as a key employment objective throughout the development of the JSB. For example, the JSB is being used to design and test the Global Strike Task Force (GSTF), and an initial focal point of this effort was the evaluation of a C2 tool, specifically, an Intelligence, Surveillance, and Reconnaissance (ISR) asset management capability, in terms of its effect on operator situational awareness (e.g., Reference 26). Future plans call for using the JSB to evaluate concepts and technologies associated with the Multi-sensor Command and Control Constellation (MC2C); the objective use of manned and unmanned air, sea, ground and space assets to collect intelligence; and the Air Operations Center (AOC) Time Critical Targeting (TCT) Cell (e.g., References 4, 27).

It is arguably possible (although not ideal) to evaluate a new aircraft system, even the pilot-vehicle interface of that system, using a simulation environment that features only systems and environmental models (e.g., if one chooses to focus on aspects other than the pilot's interactions with the technology). However, evaluation of a C2 technology or concept in the absence of accurate human representation is a much less reasonable endeavor given the emphasis in C2 on enabling effective situational awareness, decision making, communication, information flow, and information management—states and activities in which the human is central. As Bowen and his colleagues (Reference 27) note with respect to the MC2C and TCT Cell development plans, "These initiatives rely on teams of C2 operators dealing with increased information flow under reduced manpower and footprint conditions. Modeling the behavior of such teams is key to determining how advanced systems and operational concepts can best be employed to assist them."

As noted above, models used in the JSB exercise must be highly accurate and accredited (which typically implies validation has occurred). Given the potential value of CMs in contributing to the realism of the JSB and to the evaluation of C2 concepts, systems, and procedures, there appears to be a need for accredited models of the cognitive performance variety. However, in contrast with this conclusion, it does not appear that any CMs have been accredited for use within the JSB. In fact, evidence was not found by these authors to suggest that any CMs are even being considered for the accreditation process.

In the meantime, the JSB seems to have spurred a number of CM and HPM initiatives. These initiatives do not appear to be closely tied to an existing or planned JSB-based acquisition program. However, they may be of value across a range of future JSB-based acquisition activities if successful or serve as important stepping stones toward the use of CMs and HPMs in SBA. The initiatives, summarized below, include Project Gnosis (Reference 28), Cultural Modeling for Command and Control (CMC2; Reference 29), the Role Player Intelligent Controller Node (RPICN; Reference 30), and the Joint STARS Operator Surrogate Human (JOSH; References 27, 31).

3.2.1 Project Gnosis

Project Gnosis is an AFRL-sponsored research and development (R&D) project in which task network models are being developed to represent the work of staff planners within the headquarters nodes of a joint C2 network. The task modeling tool MicroSaint Sharp is being used to represent the tasks and processes performed by these staff planners as they develop products such as the Effects Tasking Order (ETO) and Air Tasking Order (ATO), with particular emphasis on the explicit representation of the available knowledge used to inform planner performance (e.g., knowledge of factors such as commander's intent, knowledge in the form of completed products made available to the next task, and knowledge in the form of expert involvement in a given task). The main objective of this effort is to support the assessment of systems, tactics, techniques, and procedures (TTPs), organizational configurations, training, and other factors in terms of their impact on the effectiveness of C2 headquarters staff processes. Gnosis is being developed as an HLA-compliant model to facilitate its future integration with combat simulation models and environments.

3.2.2 Cultural Modeling for Command and Control (CMC2)

The CMC2 effort is another recent AFRL-sponsored R&D effort, and was conducted as part of the AFRL AMBR program. This modeling effort made use of the HLA-compliant AFRL Crew Automation Requirements Testbed (CART; Reference 11), a variant of the Army Research Laboratory's IMPRINT task-network modeling tool. CART was evolved within the CMC2 effort to include micro-models of cultural differences in decision making and performance with the idea that cultural effects could then be inserted into CART-derived HPMs. Also within this effort, CMC2 was used to develop culturally enhanced HPMs, and these were used to populate a C2 structure developed within the mission-level Joint Integrated Mission Model (JIMM) simulation environment.

3.2.3 The Role Player Intelligent Controller Node (RPICN)

The RPICN is another effort made possible through the AFRL AMBR program. This effort has goals that differ from traditional CM and HPM development goals. Specifically, the RPICN goals entail developing models that serve as role-player and simulation-controller agents. As role players, the agents simulate entities in the simulation environment that are supporting an ongoing simulation exercise but that are not central participants. These agents do not necessarily perform as a human would as they typically do not require significant behavioral or cognitive performance fidelity to support the objectives of a given simulation exercise. As simulation controllers, the RPICN agents may adapt the flow of the simulation exercise or the behavior of the role-player agents. RPICN has been developed using the Operator Model Architecture (OMAR; e.g., Reference 32) and features an interface, called the *JSB Client* that allows the RPICN to interface with Air Warfare Simulation (AWSIM), a joint simulation system used to train AOC commanders, and with the JSIMS HLA Federation.

3.2.4 Joint STARS Operator Surrogate Human (JOSH)

JOSH is as an initiative that involved modeling the decision making of the Joint STARS Senior Director (SD; i.e., team leader) and certain other team tasks that influence SD decisions. JOSH was designed to process radar tracks and provide processed track data and priorities to other battlespace elements. JOSH was developed using CART (Reference 11), and was integrated with the Extended Air Defense Simulation (EADSIM), a simulation environment chosen based on its mission-level focus and a track record of successfully meeting simulation needs across a variety of USAF C2 analysis and experimentation projects.

3.2.5 Obstacles to CM and HPM Use in SBA

The status of CM and HPM development efforts associated with JSB objectives both lend hope to their future in SBA and simultaneously suggest there may be a long road ahead before CMs and HPMs are embraced by the acquisition community. In this section we describe obstacles encountered in the projects described above and that have likewise impacted other efforts to develop CMs and HPMs for SBA, beginning with a number of lessons-learned from the JOSH project. The subsequent section describes efforts that are indicative of progress toward overcoming many of these obstacles and that may lead to the acceptance of CMs and HPMs in SBA.

The overall objective of the JOSH project was to explore the use of emerging human behavior modeling techniques in military simulation systems. One of the project's most significant contributions may be the 'lessons learned' during the course of that exploration. These lessons, described by Bowen and his colleagues (Reference 27), have to do with both the modeling tool and the simulation environment used in the JOSH project, but also generalize across a range of modeling tools and simulation environments. The lessons consist of the following guidance for the successful use of CMs in simulation environments, guidance that also applies to the related domain of SBA:

- Task-network models are not suited to representing certain types of cognitive task performance. Models developed or supplemented using cognitive modeling architectures (e.g., Soar, ACT-R, and iGEN®) may avoid this problem.
- Build data exchange 'bridges' that allow CMs and HPMs developed using different tools and architectures to be interfaced to create hybrid models, and to thereby allow model developers to draw upon the strengths of different tools and architectures as well as re-use existing models regardless of their source.
- Develop better user interfaces for modeling tools.
- Develop modeling tool user interfaces and capabilities that facilitate the interfacing of models to simulation environments (such as a multi-federation object model [FOM] interfaces).
- Modify and develop simulation environments and system simulations so that they are able to accept the outputs of CMs and provide the inputs needed by CMs.
- Improve the cost effectiveness of CM development by means of model re-use.

Obstacles similar to those referenced by the above JOSH lessons-learned recently have been encountered in an ongoing effort to integrate human performance measurement agents with a U.S. Navy Manned Flight Simulator (MFS) federation that will be used in Distributed Mission Operations (DMO; Reference 33). Developers of the MFS federation agents found that the FOM

and mission scenarios—both of which were critical to agent functioning—were moving targets, constantly changing to accommodate other MFS purposes. A flexible multi-FOM HLA infrastructure layer was created to allow the agents to adapt to FOM and testbed changes; however, flexibility to accommodate a range of scenarios has yet to be accomplished. Agent developers additionally encountered difficulty obtaining human performance data needed to support agent functioning from the simulation environment. For example, a key data message that had been identified as critical to the agents at an early stage of federation development was not made available (via the HLA bus). This suggests a need for a mechanism to improve coordination between agent developers and simulation engineers, or guidance to help simulation engineers identify and support appropriate data interchange elements, or hooks, between agents and simulation systems.

The CMC2 initiative was faced with a less than tractable modeling objective. In particular, developing a model that accurately predicts the effects of culture on decision making is a challenging venture, possibly even a venture that cannot be accomplished at this juncture. Indeed, Mui et al. (Reference 29) state that inadequate data exist to develop such a model, and CMC2 features only a cultural effects micro-model framework that is still in need of parameters. CMC2 thus serves to highlight a problem that commonly interferes with the development of HPMs and CMs alike and certainly with the development of valid HPMs and CMs—a paucity of available data and empirically backed theory to inform model development processes.

Related to the issue of data paucity is the issue of CM versatility or, put another way, brittleness, and its implications for CM validation. Validation has been defined in this report as the quality and accuracy of CM simulation results when those CMs are employed in new conditions for which empirical data have not yet been collected and used to calibrate the CM. The versatility prescribed by this definition poses a significant challenge for CMs developed for large and complex synthetic military environments—a challenge that stems largely from the need to obtain and represent extensive performance and environment data in order to create a robust CM knowledge framework. Nielson and his colleagues (Reference 34, p.1) observe that "brittleness in an intelligent behavior system is a problem that arises out of ignorance (incorrect knowledge) not representation (architecture)." In other words, although the intelligent behavior system's architecture must enable the acquisition of new knowledge and the appropriate application of existing knowledge, ultimately it is knowledge that permits the intelligent model to respond to a new situation.

To avoid brittle models that cannot perform appropriately within their target performance domain when conditions change, cognitive modeling architectures such as TacAir-Soar and iGEN® have been enhanced to support the development of CMs that have knowledge about their own performance relative to objectives and performance demands, and that are able to adapt their performance based on their knowledge about changes in the environment and overarching objectives (e.g., References 34, 35). These architectural advances represent important achievements. However, they have not done away with the need to represent relevant aspects of the simulation environment and ways the CM can interact with it. Thus, there remains a requirement to develop the knowledge framework that allows the CM to interpret the simulation environment in which it is executing. In the complex military simulation domain, this requirement can translate into a time-consuming and extensive data collection and representation

effort accompanied by a temptation to implement and judge a CM as soon as it will operate in basic or standard domain conditions, even though the necessary knowledge base is only partially developed.

3.2.6 Advances Impacting Future CM and HPM Use in SBA

Despite the many challenges encountered in the course of efforts to develop CMs that may contribute to the JSB and other synthetic environments, progress is being made. This progress takes the form of advances such as:

- the identification of lessons learned to guide further progress, such as the lessons noted above;
- potential CM footholds established in the simulation engineering community, as represented by the development of immediately useful and cost-effective RPICN exercise-controller and role-player agents (Reference 30)—agents that initially may not demonstrate realistic human performance but which represent a testbed of sorts in which increasingly realistic and accurate human behavior and cognitive performance representations may be developed and evaluated within the context of the simulated battlespace;
- efforts to identify and draw together results from empirical research on different aspects of human performance in order to provide a more substantial foundation for HPM and CM development efforts (e.g., Reference 36);
- the development of multi-FOM, multi-testbed CMs (e.g., Reference 33), as recommended by Bowen et al. (Reference 27);
- cognitive modeling architecture advances to improve CM self-monitoring and adaptation capabilities (e.g., References 34, 35, 37), thereby increasing CM versatility, robustness, and validity;
- efforts to build interfaces that allow disparate CMs and HPMs to manage and control the synthetic entities in a simulation environment (e.g., References 38, 39);
- experimentation with an improved interface between CMs/HPMs and the simulation environment that features a graphical user interface (GUI) for selecting models to be used in a given simulation exercise (e.g., Reference 39); and
- concepts that involve re-engineering the acquisition process so that it may follow more of a human-centered design path (e.g., Reference 40).

Possibly to the future benefit of these and related CM advancement efforts, JSB testbeds are being developed by the AFRL Information Directorate (AFRL/IF) to facilitate the development of C2 technologies and concepts. The *JSB for Decision Support* (JSB-DS; Reference 41) is intended to provide engineers and scientists with a distributed architecture in which they can explore, develop, and evaluate C2 decision support and situation awareness visualization concepts and technologies. The initial version of JSB-DS features a simulation system, called SUPPRESSOR, that performs battlespace entity modeling at a basic procedural level, an AOC model that simulates basic TCT planning activities, a *Visualizer* tool that provides Java-based visualization and user interface functionality, and a *Common Operating Picture (COP) Knowledgebase* that consists primarily of scenario and simulation control data files.

The *JSB for Research and Development* (JSB-RD; Reference 41) is being derived directly from the JSB-DS simulation environment. The JSB-RD will serve as a testbed for C2-focused modeling-and-simulation technologies, and will be dedicated to R&D activities such as model abstraction, distributed collaborative simulation, synthetic environments, model interoperability, cognitive process modeling, human-in-the-loop simulation, and virtual prototyping.

The fact that these AFRL/IF JSB simulation systems revolve around performance models of an organization (an AOC) and entities that, in the real battlespace, would be heavily influenced by human performance capability suggests a valuable future role for CMs. AFRL/IF has, in fact, investigated the use of simulation frameworks that are more amenable to the insertion of CMs into their JSB environments. For instance, they have investigated the feasibility of replacing SUPPRESSOR, a legacy simulation system written in FORTRAN, with the newer Joint Semi-Automated Forces (JSAF) software, and additionally have identified the JIMM and VR-Forces, a commercial-off-the-shelf (COTS) CGF framework developed by MaK Technology, as alternatives worthy of investigation (e.g., Reference 42).

3.3 Joint Unmanned Combat Air Systems (J-UCAS)

The Joint Unmanned Combat Air Systems (J-UCAS) program is a joint DARPA-Air Force-Navy program to demonstrate the technical feasibility, military utility and operational value for a networked system of high performance, weaponized unmanned air vehicles to effectively and affordably prosecute missions such as Suppression of Enemy Air Defenses (SEAD), surveillance and precision strike. Boeing (X-45) and Northrop-Grumman (X-47) are leading industry teams who are developing J-UCAS platforms that can operate within a Common Operating System environment architecture. The Common Operating System is needed to support the network-centric nature of J-UCAS, the need for collaboration and synchronization, the demanding missions, and the degree of integration and interoperability needed. The Common Operating Systems is also needed to help balance the autonomy of the vehicles and the need for human monitoring and intervention. A key human consideration in the system is the ability of an operator to simultaneously monitor and control multiple lethal unmanned vehicles. Despite this fact, we have found no evidence that J-UCAS is using human performance models in its acquisition.

3.4 New Navy Ship Class – DD(X)

DD(X) is the centerpiece of a Family of Ships that will operate within the construct of the Surface Combatant Navy to deliver a vast range of warfighting capabilities that will maximize and revolutionize the combat capability of the Fleet. DD(X) is the U.S. Navy's future multi-mission surface combatant designed to deliver precision strike and fire support, dominate the littoral environment and defeat the most challenging threats. DD(X) provides the technological foundation for generations of U.S. warship classes, including surface combatants, aircraft carriers, amphibious ships and auxiliaries. DD(X) is not only developing several critical technologies and systems for the future fleet but will integrate them into a complete warfighting system that introduces powerful and effective operating principles to naval warfare, such as employment of optimal manning through human systems integration, low operations and support costs, multi-spectral signature reduction, balanced warfighting design for littoral operations,

improved quality of life, survivability, and adaptability. Equipped with state-of-the-art, network-centric information systems, DD(X) will operate seamlessly with other naval, ground, and land-based air forces. The ship's advanced command, control, and computing systems and persistent time-critical strike capability will revolutionize joint fire support and ground maneuver operations, while also providing strike group or joint task force commanders the freedom to direct aircraft to other high-value targets of opportunity, as appropriate.

The DD(X) program completed a successful system-level preliminary design review in March 2004 and separate preliminary design reviews of the Engineering Development Models over the past two years. A team, led by Northrop Grumman Ship Systems (prime contractor) and Raytheon (weapon and electronic systems integrator), is leading the DD(X) design effort. The Navy plans to award detail design and construction for the lead ship of the class in March 2005. The lead ship is scheduled for fleet delivery in 2011 and to enter service in 2013.

The DD(X) program is focused on developing 11 key Engineering Development Models (EDM) to demonstrate technologies critical to future warships. The 11 EDMs include electric drive and integrated power management systems; multi-function and volume search radar suites; the Advanced Gun System; and new hull design emphasizing efficiency at 30-knots sustained speed, mission payload growth capacity and stealth. The program is using extensive modeling and simulation to support EDM development and system testing and integration activities. Task network models are being used to assess key human interface issues and assess operator and maintainer workload, but we have found no evidence that cognitive models are being used for DD(X) acquisition.

3.5 Army Future Combat System (FCS)

The US Army's Future Combat Systems (FCS) is a key program in the Army's transformation effort. The FCS will provide a system of systems with an extensive distributed, networked C4ISR capability primarily for direct fire, indirect fire, air defense, and troop transport. Boeing and SAIC together comprise what's termed the Lead Systems Integrator (LSI) team, which functions much like a general contractor in overseeing and ensuring all program objectives are met and continuously soliciting expertise from around the world. The FCS LSI Team is partnering with Industry and the Government to provide the Army with the most effective and best value FCS solution for fielding in 2010.

The FCS acquisition program is using the SMART approach (see above) which uses modeling and simulation to facilitate more effective and efficient collaboration among FCS functional areas and team members, including government, military and industry. (See Reference 43) The FCS program is using a task network modeling approach with IMPRINT to assist in system design while addressing key FCS functions, such as communicating information, scanning for targets, identifying targets, shooting targets, commanding troops, driving vehicles and maintaining situational awareness. For example, IMPRINT has been used to examine the workload impact of various combinations for two and three person crews for FCS combat vehicles by spreading the crew functions across the commander, driver and gunner. IMPRINT is

a dynamic, stochastic discrete event network modeling tool designed to help assess the interaction of soldier and system performance throughout the system lifecycle--from concept and design through field testing and system upgrades. In the course of our investigation, we have found no evidence of the use of cognitive models in the FCS program.

3.6 Army SMART

The Army's current response to the various DoD SBA policy directives is realized through the Simulation and Modeling for Acquisition, Requirements and Training (SMART) approach, adopted in 1997. Von Holle (Reference 4) notes that the SMART approach "appears to be the most vigorous and institutionalized effort among the services." The Army Model and Simulation Executive Council (AMSEC) is the institutional proponent for the SMART concept, acting through the Army Model and Simulation Office (AMSO) (Reference 44). AMSO maintains and publishes the SMART Planning Guidelines (Reference 45), validates simulation support planning, shares lessons learned, and promotes and supports the application of the SMART approach by organizing annual conferences and by providing various educational materials via the Internet. The SMARTeam (Reference 46), a collaborative effort with several other organizations including the U.S. Army Program Executive Office for Simulation, Training, & Instrumentation (PEO STRI), provides direct support to project teams for both simulation support planning and after action reviews of plan implementations.

As its name suggests, a key part of the SMART approach is the collaborative use of modeling and simulation (M&S) across the requirements, acquisition, and training communities. A driving consideration is the direct sharing of costs associated with modeling and simulation, but SMART is also an attempt to move the culture away from isolated communities of interest through shared responsibility for and use of M&S technologies. If system requirements, engineering solutions, prototyping, testing, manufacturing, system support, and training can be addressed in some common domain, it is believed that the resulting system will be better as well as cheaper. That common domain is modeling and simulation, with related M&S technologies being used throughout the system acquisition process. In a sense, SMART is as much about fostering and supporting collaboration as it is about modeling and simulation. This can be seen in the current SMART planning guidelines which state that the "fundamental elements" of SMART are:

- Collaborative Environments (CE)
- Distributed Product Descriptions (DPD)
- Modeling and Simulation Standards and Reuse

The collaborative environments envisioned include the interoperable tools, information sources (SMEs as well as databases), and process models required to pursue the acquisition process. A CE will provide the workspace where the collaborating stakeholders can identify and resolve the tradeoffs associated with a system throughout the acquisition process. Although the SMART guidelines assert that such environments are critical, it is noted that these infrastructure technologies are still evolving. No single CE (product) is explicitly mandated by the guidelines. A recent SMART Lessons Learned case study (Reference 47) describes a situation where contractors set up their own CEs because they did not trust the security of the system the Army could provide. The resulting barriers to information sharing obviously fly in the face of

SMART's goals for collaboration. Although each acquisition program is currently expected to resolve the CE issue on its own, efforts are underway to provide a common CE across acquisition programs.

The Advanced Collaborative Environment (ACE) is one of the evolving technologies which can support the implementation of the SMART approach. ACE is a Web-accessible information management system, integrating M&S and other information sources into a single collaborative environment for use across all acquisition communities, through all phases of the acquisition process. ACE provides integrated support tools, such as process control via workflows and 2D and 3D visualization for prototyping, as well as network connectivity to reach data stored in multiple internal/external repositories. ACE is implemented using commercial-off-the-shelf (COTS) technologies, and relies on commercial and defacto standards to provide integration with various common tools and data sources. ACE is currently being implemented for and used as the collaborative environment of the Army's Future Combat System (FCS) Program (a system of systems, including a variety of ground systems, air vehicles, sensors, and network links), giving government and contractor personnel access to information sources critical to the FCS acquisition process. Earlier this year, the FCS program manager was presented an Army Knowledge Award for ACE, on the grounds that ACE had already reduced the FCS program timeline. ACE is clearly moving forward and is already a prime candidate for fulfillment of SMART's CE requirement, although it is not yet a mandated solution.

The Distributed Product Description (the second fundamental element of SMART as presented in the guidelines) speaks to the organization of product information, including such elements as product cost data, engineering data, test data, and business processes, as well as models and simulations of a product's behavior or performance. This requirement addresses the need of users to view a wide range of different types of product information through a unified representation. Although information elements may be distributed, they must be interconnected for purposes of access. Like the CE requirement, the DPD requirement does not mandate the use of a specific tool or systems. Currently ACE appears to satisfy the DPD requirement and will undoubtedly contribute to ACE's widespread use and perhaps ultimately to its mandated use under SMART.

Although the CE and DPD elements described in general terms in the SMART guidelines are rapidly becoming a reality, at least partially as a result of their roots in COTS technologies and commercial best practices, the need for "modeling and simulation standards and reuse" is somewhat further behind the curve. As in the CE and DPD cases, no specific product solution is advocated, but here several different M&S repositories which "exist to facilitate the reuse and standardization of M&S" are identified:

- The Defense Intelligence Modeling and Simulation Resource Repository (DIMSRR) for validated threat M&S products.
- The Army Standards Repository System (ASTARS), used to store approved Army standard algorithms and models, standards nominations with standards requirements documents (SRDs), and supporting documents.
- The Army Model and Simulation Resource Repository (MSRR), part of a larger DoD effort, a collection which includes models, simulations, object models, conceptual models of

the mission space (CMMS), algorithms, instance databases, data sets, data standardization and administration products, documents, tools and utilities.

The SMART guidelines user is here directed to what seem to be competing M&S reuse repositories and is then also directed to the U.S. Army Materiel Systems Analysis Activity (AMSAA). AMSAA is the Army's center for item-level performance analysis and certified data as well as the "publication of Joint Munitions Effectiveness Manuals (JMEM), single source documents for modelers, materiel developers and strategic and operational planners at all levels." In the area of M&S standards and reuse, AMSO is not promoting a centralized vision and is apparently not active in developing such a vision. This is particularly noticeable in the VV&A section of the guidelines. The presentation consists of little more than definitions and a clarification of proponent and developer roles in V&V (the M&S proponent is ultimately responsible for V&V) based on AR 5-11 (Reference 48).

The clearest requirement fostered by AMSO through the planning guidelines is the Simulation Support Plan (SSP). The SSP is the key document in the implementation of the SMART approach. As different uses of the same or similar models, possibly by different communities, may arise in different phases of the same acquisition process, it is critical that model usage be carefully planned. While it may be difficult to reuse a model developed years earlier under a different program for a different purpose, it should be possible to anticipate this kind of reuse within a program where the different communities have related needs. For example, the same threat models may be required for early analysis of alternatives (AoA) as well as the later construction of training components. The SSP is critical for the SMART approach because it is directly related to the collaborative focus of SMART, bridging issues of time as well as community. The program manager uses the SSP to frame a SMART acquisition strategy which will lead to the expected benefits over the course of the acquisition process. Within the SSP context, M&S products are identified and assessed for their ability to generate information and support decisions. Specifically, the guidelines state that:

> *At a minimum, a simulation support plan accurately records M&S activities undertaken in support of materiel requirements determination or program acquisition. The SSP should also discuss coordination with other organizations and planned future M&S activities. Simulation support planners must provide rationale for decisions to employ M&S. This rationale could be discussed as a part of the crosswalk that links program requirements and/or issues with planned use of specific models and simulations. (p. 71)*

In general, all types of models and simulations are equivalent in the SSP perspective. Although the distinction between constructive, virtual, and live simulations is recognized, as are various M&S uses (combat development, design and engineering, logistics and support, test and evaluation, training, and life cycle management), the SSP treats all models and simulations as equivalent for the purposes of planning. It is assumed that an SBA process which incorporates human performance models and cognitive models should be planned in the same way (with the same issues considered) as an SBA process which incorporates only algorithms and mathematical models on the order of threat vulnerability assessment tools.

On their Web site, AMSO has made two approved SSPs available as examples for SSP developers. A close examination of one of these documents gives some insight into why HPMs and CMs are not singled out for specific mention. The Modernized Longbow Apache Block III program SSP (Reference 49) lists 89 "models that have been or may be involved in the Block III program." Of the 89 models listed, 10 are essentially constructive simulations involving something which might be termed human performance modeling, typically unit level decisions based on an encoding of doctrine. 15 of the 89 are either models explicitly intended to support human-in-the-loop (HITL) studies or are intended to function as (more or less immersive) trainers which may double in a HITL role. Only a single model in the list of 89 is obviously a human performance model and specifically a cognitive model (and included Pew and Mavor's framework comparisons, Reference 1) – the Man-machine Integration Design & Analysis System (MIDAS). Although the SSP describes the general utility of MIDAS in addressing crew station design issues, the only specific applications of MIDAS noted for the Block III program are a "partial demonstration with the AH-64A" and the use of MIDAS' imbedded Binocular Vision Model "partial simulation for the AH-64D Longbow Apache." No future use of MIDAS is projected. It is perhaps worth noting that although the second SSP provided by AMSO as an example (the Warfighter Information Network – Tactical SSP, Reference 50) includes doctrine-based constructive simulations and HITL support, no use of cognitive models is indicated.

Although SMART may indeed be the most "institutionalized" effort to implement SBA in the military, the possibility that human performance models may require special attention in the SBA process (e.g. the acquisition of data for cognitive model validation may need to be explicitly addressed) does not seem to be on AMSEC's/AMSO's radar. The complexity of the SBA issues already addressed by SMART, the focus on collaboration issues (perhaps at the expense of modeling and simulation employment issues such as standards, validation processes, etc.), and the highly limited use of cognitive models in current Army acquisition practice, are likely to keep the need for special attention to HPMs and CMs from emerging as a significant concern in the near future.

4. HPMs, CMs and Alternatives in the Current Acquisition Process

A variety of tools and methods are currently available for the construction of HPMs in complex system and task environments (e.g., Reference 1). Among other purposes, these models are being considered for use in SBA, providing tools for the interim evaluation of system performance at many intermediate stages in the acquisition process. Such HPM models can be used to represent human components of the system to be acquired, and/or human elements of the environment in which such a system would be employed. The various modeling tools offer somewhat different representations of human behavior and also offer different kinds and amounts of evidence for the validity of the tool and technique. There are numerous cases where models have been constructed and used to generate performance predictions, followed by collection of real human performance data for the model context and evaluation of the correspondence between the predicted and empirical data.

But there are very few cases where multiple modeling frameworks have been applied to a complex system application and comparatively evaluated relative to empirical ground truth (with the Air Force AMBR program (Reference 12) representing the principal such endeavor). To our knowledge, there are no cases where such comparative HPM evaluation has addressed the specific requirements of SBA. At the same time, we recognize that system performance data are frequently collected in the T&E stage of acquisition, which occurs much later in the development process and typically involves an entirely different team than that which conducted the simulation-based evaluation earlier in acquisition. Because the necessary connections between simulation-based evaluations and empirical T&E evaluations are not always forged at the outset of acquisition, it should not be surprising that comparisons of these two sources of system evaluation data are not typically performed or easily accomplished.

Several recent survey efforts have identified and documented the salient characteristics of methods and models applicable to SBA (e.g., References 1, 2, 51, 52, 53), greatly facilitating our task to identify these candidates for the present investigation. In addition to identifying the aspects of human performance that are addressed by each of the HPM candidates, we must also determine how these techniques are typically used and identify the components and issues for validation. Do they employ libraries of micro-models or performance moderator functions? Do they offer any guidance, loose or structured, for development of model applications? Are they usable by typical human factors analysts with modest amounts of special training, or do they require extensive special training to the point where they are only used by the tool developers?

In addition to identifying and evaluating the principal human performance modeling methods and tools, it is also necessary to identify the principal alternative analytic techniques that do not employ human performance models but still address the same kinds of design evaluation questions for military systems acquisition. After all, SBA is a fairly new concept and all of the military services have acquired all of their systems through structured acquisition processes for many decades. Although crude, unaided judgment and intuition have certainly been used in many cases, a variety of ad hoc techniques have been developed for other cases, and a few general analytic techniques to refine human judgment have also been offered. One fairly sophisticated technique for using "anchoring and adjusting" processes to extrapolate from legacy system performance characteristics to proposed new system performance characteristics was

offered by the Navy's HARDMAN methodology, which was itself subjected to a fairly elaborate validation study (Reference 54). The essence of this technique is to identify a legacy system or component with know performance characteristics and develop analytically guided estimates of how those performance characteristics should be expected to change as a consequence of the design differences between the new system and the legacy system. Another more recent and more tractable technique was developed in the course of the Navy's Advanced Technology Crew Station (ATCS) program in the form of a Performance Metrics Methodology (Reference 55) that employs a variant of the Quality Function Deployment (QFD) management science technique to generate analytic prediction estimates for the impact of new system design features on human-system performance. This technique requires the user to develop analytic identifications of measures of performance (MOPs) for relevant human system performance criteria and separate measures of effectiveness (MOEs) for the overall system, along with estimates of relative priorities and interconnections between the two types of measure and some selective use of HITL experimentation to resolve ambiguities. But probably the most widely used non-HPM technique for SBA is VE-HITL simulation, which allows developers to obtain empirical data about human performance and usability with a new system design. It is important for us to include such alternative techniques that involve no recourse to human modeling in model validation studies so that we can identify some baseline of prediction performance to which we can compare model results.

Because the non-HPM-based techniques will generally be much quicker and cheaper to implement than the model-based techniques, it is also important to identify domains and requirements for which HITL T&E is not feasible, and to determine under what circumstances model predictions are significantly superior to non-model predictions. It is conceivable that we could sometimes find that models do an adequate job of predicting human performance and supporting SBA decisions, but that unaided expert judgment or expert judgment aided by a simple non-model tool might do as good a job at a much lower cost. Validation in this sense must take into consideration all plausible available alternatives for doing the same job.

Although we did not want to divert our investigation to studying the alternatives to HPMs in too much detail, it is important to recognize the implications that they have for CM validation. Any practical validation of an HPM for SBA purposes must also consider the other alternatives for SBA design evaluation decisions that establish the criteria for quality and confidence in the HPM analysis option. The most important factor here is the availability of VE-HITL evaluation of system designs. Through the early stages of this project we have considered the VE-HITL alternative as a distinct separate and competing choice from that of direct HPM (and especially CM) use for the primary HITL roles, but we also still need to address validation issues associated with HPM/CM usage for representation of peripheral roles (teammates and adversaries) in VE-HITL applications. It is also noteworthy that the relevance of these VE-HITL cases is likely to be more central because of the expectation that VE experiments will probably be the primary mode for design T&E (DT&E) activities.

Thus, the principal options for design testing seem to be the following:

a) Subjective SME judgment without guidance – mainly just asking the SMEs what they like and don't like based on descriptions or mock-ups of the design – this is clearly a cost-effective technique and should almost always be pursued as a first option, but with

the recognition that SME tend to exhibit a variety of biases along with unreliable insight into the structure and constraints of their own behavior;

b) Subjective SME judgment with guidance – using analytic structures such as Navy HARDMAN (anchoring and adjusting), the QFD Performance Metrics technique, or task network workload estimation – these techniques help to resolve some of the uncertainties and biases of unstructured SME judgments, but the accuracy and reliability of the results are still uncertain;
c) VE-HITL experiments without CMs – using VE testbeds to simulate the system and environment but requiring real people to play all relevant human roles – this tends to be considered as the most reliable option, though it can be very costly and difficult to control when many peripheral role players (teammates and adversaries) must be involved;
d) VE-HITL experiments with CMs for peripheral roles – using VE testbeds with CMs playing some peripheral roles for teammates and/or adversaries, but with real people playing the primary roles – this approach can reduce the costs and increase the experimenter control for the role players, though it can also introduce questions about the contributions of invalidated performance characteristics of the CM role players;
e) Constructive simulation experiments – using possibly the same VE simulation testbed but with CMs playing all relevant human roles – this approach just expands the concern about CM validation to the primary role along with all of the peripheral roles.

Viewed in this fashion, it is apparent that the last option of complete constructive simulation is not necessarily very distinct from the prior VE-HITL+CM stage.

Although not true of all HPMs in general, it appears that CMs are typically developed to integrate into a host simulation environment for which the maximum fidelity VE simulation that is available is the prime candidate. Thus, it seems that the VE testbed is really a prerequisite to the employment of any CM. This is necessary as long as the CM requires the kind of rich knowledge representation that can only be generated by a high-end simulation. And even if the VE testbed is not developed separately from the CM, the CM will still require the establishment of a system-and-environment simulation that easily could be configured as a VE testbed. Thus, it seems that the constructive CM option will always entail the ready availability of some kind of VE-HITL option as well, so there is not likely to be any clear cost saving in a plan to employ a constructive CM evaluation in order to avoid the costs of a VE-HITL evaluation.

The distinction between calibration and validation then comes into play to further clarify the appropriate relationship of these options. Our view is that validation refers to the quality of and appropriate confidence in CM simulation results when we build and execute the simulations for new conditions (new system designs, missions, scenarios, user characteristics, etc.) for which we have not yet collected (or at least seen) empirical data. The actual comparison of model and empirical data also constitutes a relevant measure of validity as long as the model is not adjusted after the data becomes available. In principle, we want to develop measures of CM validity so that we don't need to collect empirical data and make these comparisons in every case. After we adjust the CM (structurally or parametrically) to achieve a best fit to the empirical data, then we have a calibration point. The act of obtaining a large collection of calibration points will not necessarily provide a useful indication of model validity unless we know how well the model fits the data in each case before any a posteriori adjustments were made. It is also noteworthy that the ability to fit the model to the empirical data will also constitute some measure of model

validity – certainly in the case of the AMBR category learning problem; it was not obvious that all of the modeling efforts would be able to achieve a very good fit.

Our most recent view of the relevance of VE-HITL experiments and CM calibration efforts to validation is that VE-HITL is the primary medium for DT&E for SBA programs. Ideally, every VE-HITL experiment (assuming no peripheral roles or just peripheral CMs) should provide an opportunity for a CM calibration point if a constructive CM simulation can be built for the experimental condition. If we could create the discipline to compare CM results to the experiment before any adjustments, then we could thereby develop a basis for projecting model validity to new data points associated with variations in design, mission, scenarios, users, etc. This is just a kind of curve fitting or response surface methodology in a complex multidimensional space. But there is an important special constraint in the fact that data points are typically very few and difficult to come by.

So the emerging view here is that CM validation will be developed primarily through the integrated use of VE simulations and CMs that use the same underlying simulation engines. The VE-HITL experiments will provide the data to calibrate the CM and to develop a basis for validation of the CM. Of course, it is then necessary to demonstrate that there are useful applications of such validated CMs that cannot be more cost-effectively accomplished using just the VE-HITL experimentation. Such applications are extremely important to identify and explore in order to motivate any serious use of CMs in this fashion.

One potential advantage of CMs that has been noted in this regard (Reference 56) is the value of CM transparency for diagnosing performance deficiencies in terms of component performance resources or characteristics that can then suggest appropriate redesign remedies. For example, a CM might indicate that an excessive response time to a critical situation could be localized to perception and situation assessment rather than to evaluation of action alternatives, which could then focus redesign efforts on information fusion and display rather than organization of information about response options.

There are also some types of evaluation which, although experimental in concept, are generally impossible to run as experiments with human participants because of the difficulty of generating all of the required data points – sometimes because there are too many data points that are required and sometimes because the specific types of required data points are difficult or impossible to achieve. A case in point is sensitivity analysis which maps performance to systematic variations in various design (especially human interface) or user or environmental factors. If data for several human participants are required for each parameter value of the factor, with each data collection session being lengthy and costly, and if many parameter values must be included in the analysis, then it will be prohibitively costly to run such a sensitivity analysis. This is a serious problem if the sensitivity analysis is the best or only way to determine the right parameter settings for an interface design. Another case is the population accommodation analysis, where it may be very difficult to obtain real people to represent all of the segments of the population that must be accommodated. This is particularly problematic when multiple dimensions of cognitive variation of the user population must be simultaneously addressed.

5. Validation of Simulations and Cognitive Models

We will focus in this section on issues of validation of HPMs with a particular focus on the case of CMs. We will first introduce the notion of "application validity" as background to validation in the SBA context. We will then consider some general validation issues, from several perspectives relative to SBA. It is noteworthy that, although our primary focus will be on the case of CMs, most observations will apply equally to most other HPMs.

5.1 Application Validity and SBA

The Agent-based Modeling and Behavioral Representation (AMBR) program, a multi-year effort sponsored by the Air-Force Research Laboratory (AFRL), was undertaken to advance the state of the art in cognitive simulation and human behavioral representation. Four modeling teams participated in a "fly-off", comparing results from models developed under different cognitive architectures (ACT-R, COGNET/iGEN®, DCOG, and EASE, the latter an ACT-R/Soar/EPIC hybrid). Two separate experiments were performed, addressing the modeling of different "operator" tasks. The Experiment 1 task was a simplified air traffic control (ATC) task, focusing modeling efforts on multi-tasking. The ATC task can be broken down into three high-level categories of (frequently interleaved) actions, related to responding to incoming aircraft, outgoing aircraft, and requests for speed changes. Experiment 2 added a requirement for modeling concept learning in the context of the ATC task. Specifically, appropriate responses to aircraft requests for altitude changes needed to be learned by associating feedback concerning the correctness of responses with three feature dimensions associate with each aircraft. The ATC simulation environment was constructed to allow either a human or a model to play the operator role and interact with the ATC GUI to perform the requisite task actions. In both Experiment 1 and Experiment 2, task performance data were collected from all four models as well as human subjects. The AMBR data provide a unique opportunity for the comparison of real and simulated human cognition and behavior, and have resulted in a series of publications (see Reference 57). Two papers are of particular interest here, as they use the AMBR results to address issues of cognitive model validation (see References 13, 58).

5.1.1 Construct validity

In order to discuss the results of the AMBR experiments, Campbell and Bolton (Reference 58) introduce the term "construct validity." Drawn from the literature of experimental psychology, the concept is applied here to the validity of a model as a representation of human cognition and behavior. They note that a construct valid Human Behavioral Representation (HBR) "would be one in which the knowledge base and behavior engine implemented in the model correspond to the knowledge structures and cognitive and psychomotor processes of the person or people being modeled." These processes cannot necessarily be measured or studied directly, but rather are understood through inferences, accumulated as evidence over time. Validity here is not "all-or-nothing," although specific communities may establish thresholds of acceptability. Campbell and Bolton proceed to review the qualitative and quantitative methods for assessing validity, and note their employment in AMBR, as described below.

5.1.2 Qualitative assessment of construct validity

The most common form of qualitative assessment in use is based on the judgment of subject matter experts (SMEs), experts in the task area being modeled. The alignment of model content and behavior with common experience is generally termed "face validity." Due to the limitations of human observers, subjective SME judgments are not considered strong evidence in the academic community. In any case, this type of assessment was not possible for AMBR. As each human subject and modeling team developed their own strategies for task performance (task performance was not trained), there were no SMEs to make judgments against common experience. A similar situation frequently arises in the real world of SBA where no similar task or task environment may have existed previously and modelers are faced with modeling what Young calls "conjectured behavior" (Reference 13).

Campbell and Bolton point out that there is an alternative source of SMEs who can assess construct validity – experts in psychological theory, or alternatively, experts in the knowledge structures and cognitive and psychomotor processes of interest. Such individuals can presumably judge if a model's knowledge base and behavior engine are consistent with current theory. AMBR convened a panel of psychologists and cognitive scientists to make just such judgments. Unfortunately, both Campbell and Bolton, and Young agree that in the analysis of the AMBR models, being "consistent with theory" or "psychologically valid" was largely in the eye of the beholder. Young notes that "Scientists with different theoretical orientations tend to have different perspectives on what is a psychologically valid implementation, and what is not." In the end the panel generally agreed that all of the models were "reasonably" psychologically valid, a position which suggests that psychologist SMEs can provide no stronger evidence for construct validity than task expert SMEs.

5.1.3 Quantitative assessment of construct validity

Although the concept is drawn from experimental psychology, Campbell and Bolton note that quantitative assessment of construct validity cannot be approached like hypothesis testing. In hypothesis testing we typically want to show that two samples are unlikely to have been drawn from the same population, rejecting the null hypothesis. In assessment of validity we want to show that two samples (human and model) could have been drawn from the same population. In hypothesis testing terms this would be equivalent to proving the null hypothesis true, an impossible situation. In Campbell and Bolton's words, "the lack of statistical significance between a model's predictions and a set of empirical data does not constitute evidence for the validity of the model"

The principal alternative to inappropriate hypothesis testing procedures is to focus on goodness-of-fit between a model's predictions and empirical data. Campbell and Bolton note that assessment along two dimensions is preferable – exact match and trend consistency. Both of these techniques were used in assessing the AMBR data, the G2 statistic (deviance) or Sum of the Squared Error (SSE) to assess exact match, and separate analyses of variance on human and model data, checked for the same pattern of significant results to assess trend consistency. Overall, the AMBR models all fit the human subject data quite well. As Campbell and Bolton point out, however, a close fit can derive from either the fact that the model is construct valid, or

that the model is powerful enough to fit (or overfit) any set of data. At the same time, these authors identify three ways to strengthen the goodness-of-fit evidence for construct validity:

1. compare fits across different models,
2. assess fit to multiple data streams, and
3. assess fit of model to data not seen.

The AMBR experiments employed all of these techniques. Goodness-of-fit was compared across all four models in all cases. Fits varied in detail, and occasionally a single model would exhibit a poorer fit than other models for a specific measure, suggesting lower construct validity in that area. The model and human data included a diverse set of performance measures in both experiments, over which all models generally produced good fits, indicating considerable strength in construct validity. The more complex the pattern of data fit by a single model, the more confidence one can have in the validity assessment. Experiment 2 offered an opportunity to assess the fit of AMBR models to data not seen. The experiment included a "transfer task" in which human subjects and models were asked to categorize stimuli with values that they had not previously seen. All four models failed to predict the human behavior observed. This result gives some indication of the brittleness of models, which otherwise seem relatively valid, when faced with novel conditions. Judging the limits of model reuse clearly pose a significant challenge.

Both Campbell and Bolton and Young note that the goodness-of-fit analyses done in AMBR were based on aggregate data and not individual data. The model with the best fit to aggregated data may not, in fact, be construct valid. The aggregated data need not represent any actual human behavior – performance results averaged across disparate strategies may represent no single strategy. Although more costly to provide, goodness-of-fit to individual data is often of particular interest in military contexts, as in support of system design where performance at the extremes of the behavioral continuum may be critical.

5.1.4 Application validity

The Defense Modeling and Simulation Office (DMSO) defines validation as "...the degree to which a model or simulation is a faithful representation of the real world from the perspective of the intended uses of that model or simulation." (Reference 59) Campbell and Bolton introduce the term "application validity" to capture DMSO's notion of validity for "intended use." While construct validity is of special interest to the psychological community as related to theory development, the military is interested in improving military capabilities. Using models in the improvement process has implications for what constitutes validity.

Campbell and Bolton use the example of training systems which use human performance or cognitive models (e.g., as adversaries). Such systems many not require construct valid models. The true test of validity in this context is whether or not the model-based training improves trainee performance. If performance improved, the model could be considered application valid. Models used in decision support systems (DDSs) could be evaluated in a similar manner. That is, if an operator using a model-based DDS exhibits a decrease in the number of incorrect decisions or an increase in the speed of decision making, the model could be considered application valid.

As Campbell and Bolton note, the situation is quite different with models used in SBA. Although one can imagine actually assessing model validity based on the success or failure of a larger training or DDS system, no one is going to use the success or failure of an acquisition process to judge the validity of a model. Certainly not intentionally. This suggests that the assessment of application validity for SBA may of necessity be equivalent to the assessment of construct validity seen in AMBR (the AMBR models have no "use" outside of the exploration of modeling issues and therefore cannot have application validity in any real sense). Campbell and Bolton recommend the following:

> *A partial solution to this problem is to divide the acquisition process into phases and determine whether or not the use of an HBR supports the team's goals in each stage. For example, the use of an HBR should be able to complement the more traditional practice of building a prototype and conducting human-in-the-loop studies to assess a candidate design. Thus, one requirement for the application validity of an HBR in this context is that the results of a simulation with an HBR must be comparable (to some degree of precision) to the results that would have been obtained by building a prototype and conducting human-in-the-loop studies. (It should be noted that, in some cases an accurate rank ordering of candidate designs based on mission outcomes will suffice.)*

Taking the notion of dividing the acquisition process a step further, SBA "use" may actually constitute a large number of different uses and require a similarly large number of different approaches to application validity. Campbell and Bolton themselves note several other model uses in SBA, including the identification of design aspects which could be improved and the rapid exploration of a much larger region of the possible design space than could ever be addressed through prototypes. Having noted that "the intended use of the model strongly influences the degree of validation that is required," Young suggests that "models used for training purposes can frequently be face validated, whereas models used for engineering and simulation-based acquisition need to be more fully validated." If SBA actually involves a series of different uses, it is possible that they range from those requiring no more than face validation to those which require very stringent validation.

If, for example, models with limited validity could expose pitfalls in a large design space during early concept development phases, and be of use further down the line for finer grained analyses, these models could continue to be developed throughout a longer term acquisition process. Over the course of the process it might be possible to bootstrap model validity, moving from uses requiring less fully validated models to those requiring more fully validated models. This transition would be very similar to the more general process described above, with a body of validation evidence growing over time, ultimately leading to community acceptance, here within the scope of a single acquisition process.

It should be noted that even though the AMBR research involved no assessment of application validity, this discussion of model validation in SBA (as opposed to validation in training or DDS development) suggests that the example is still pertinent. In Campbell and Bolton's words: "While there are limitations to the applicability of the psychological notion of theoretical construct validity to specific, individual HBRs, we argued that the types of evidence

associated with establishing construct validity are potentially quite relevant and useful to the applied military M&S community."

5.2 General issues in CM validation

For each of the roles that we can envision for CMs in SBA, the question of primary interest is "what is the best and most affordable approach (or technique or methodology or technology or design feature) to employ in order to address each of the functional requirements that are derived from the overall system requirement. Thus, we are not primarily interested in how good the predictions of a simulation (or class of simulations) might be in any absolute sense, but rather in how good the simulation outputs are in comparison to the other candidate sources of design advice that might be available. Other candidate sources of design advice might include, as discussed above, subjective reports from SMEs, empirical results from various forms of design implementation and test (e.g., various mixes across all of the major human/system/environment components between virtual and real implementation), as well as all of the alternative analytical model formulations.

After we decide that the simulation (with CM) is indeed the best method to use to answer the immediate design question, then we still have reason to be concerned about how good such answers might be. Just because they happen to be the most accurate, reliable, or cost-effective answers that are available doesn't mean that there isn't any residual uncertainty and risk. Knowledge of the uncertainty or other variability around each of the simulation predictions can be valuable in guiding the appropriate focus for design refinement and design evaluation in SBA.

Thus, there are two primary reasons for SBA to obtain information about the likely quality of model outputs (i.e., be interested in model validation):

(1) to aid in deciding whether to use a particular simulation or some other technique to address a design evaluation requirement, and

(2) to qualify the various measures of quality for simulation predictions (i.e., accuracy, reliability, uncertainty, variability, bias, etc.).

Of these, it would seem that the first will generally deserve highest priority, except in (presumably rare) cases where it could be shown that differences in predictive quality across candidate techniques was smaller than the uncertain variability in predictions from each technique by itself.

When we talk about validation of models, especially CMs, in the context of SBA, we should be principally concerned with these two purposes for gathering information about simulation candidates. Because of the primary focus on supporting the decisions of what techniques to use for what purposes in SBA, the information that is developed and provided through validation processes should be framed along the line of cost-effectiveness and return-on-investment (ROI). So we need to look at cost measures in conjunction with effectiveness measures. On the cost side, we have the obvious factors of level-of-effort (LOE) and time-to-complete, along with some more complex factors associated with longer term investments and organizational commitments that are required to support each approach. The costs of developing and maintaining testbed capabilities, expertise and facilities for CM and simulation in general and even for ensuring availability of appropriate SMEs should all be considered. It may also be

appropriate to consider amortization of developments of tools, techniques, and even personnel capabilities that are expected to have finite, identifiable spans of application.

Further examination of the calibration-validation continuum is warranted in the context of all of the above observations. In an idealized sense, we can view validation as concerning the extrapolations and interpolations relative to the fixed calibration data points in the multidimensional space of potential simulation applications. In accomplishing validation, we would like to speculate about the range of errors that should be expected in terms of the differences between simulation outputs/predictions and corresponding measurements in some sort of empirical reality. At the calibration points, we actually collect the empirical data and adjust the model to optimize the fit, thus producing a known residual error. But when we adapt these 'calibrated' simulations to new places in the application space (slightly different system, environment, scenario, and human operator specs), we can only speculate about the likely size of such errors. Clearly, the quality of any validation should improve with the expansion (and appropriate distribution across the application space) of the calibration points that represent the anchor points for extrapolation and interpolation.

From the SBA perspective of design engineering, the concept of model validation must be understood as pertaining to the advisability and expected effectiveness of reusing a model developed in some other context for an immediate context and purpose. This contrasts significantly with the more familiar notion from pure/theoretical science that model validation is a confirmation that a model representation is the "best possible" and "most efficient" description of a process or phenomenon of interest in terms of identification of relevant factors, knowledge, and component processes and the dynamics of their variation and interaction. From the perspective of pure science, Newtonian mechanics was invalidated by the emergence of relativistic mechanics. Nevertheless, from the perspective of most terrestrial engineering projects, Newtonian mechanics is sufficiently valid for reliable application in design and construction of bridges, buildings, motor vehicles, and most other human-scale engineered objects. But then other models are typically required when the engineer moves out of the human-scale realm into areas such as space exploration or nanotechnology.

A major distinction between HPM models for SBA and physical system engineering models is the recognized need for HPMs to be sensitive to many complex aspects of application context as opposed to the representational austerity of physical models. This is true particularly because the primary human roles in advanced technology systems are to mediate the processing of information related to a broad range of context factors. And it is many of the context factors that occur in model uses in SBA that further complicate the model validation process.

Model validation is much more than the simple comparison of model predictions with empirical data and the binary determination that the model is or is not valid. It is appropriate to view validation on a continuum of processes. At one end of the continuum is model calibration where we use the discrepancies between actual model predictions and empirical data to adjust parametric or structural aspects of the model in order to improve the correspondence for a subsequent execution of the same model. In support of calibration, diagnostic functions are necessary to determine what levels or aspects of the model are responsible for any observed discrepancies. Attribution of distinct aspects of the prediction-observation discrepancies to

distinct model facets can serve to facilitate both the calibration of the appropriate facet and the judgment of which facets are working adequately without further adjustment. Without the ability to attribute observed prediction-observation discrepancies to specific aspects of the levels and elements of the HPM processes and tools (i.e., model architecture, knowledge elicitation process, model components, parameter estimation, etc.), it would be very difficult to accomplish incremental improvements on any initial model application. Since major model applications can be fairly costly to undertake, it is essential to be able to have effective techniques for making such incremental refinements in a calibration stage of implementation of the model applications.

At the other end of the continuum, we have fundamental inquiry regarding the inherent value of different modeling frameworks, paradigms, and philosophies. While our methodology acknowledges the need to make such critical judgments, the social and temporal dimensions of such inquiry is well beyond the scope of our present effort. However, between this end-point and simple calibration lies the realm of practical model validations in various forms, depending on the degree of generality and scope of the model being validated. At the least general end of the continuum, the validation of specific model configurations in detailed contexts degenerates into the case of simply adjusting model parameters to achieve a valid model "variant" in a given instance of use (i.e., calibration). Moving toward the more general side of the continuum, increasingly broader classes of models are validated in contexts that are correspondingly more general. For example, a general architecture such as SOAR or iGEN® might be validated for a class of applications such as pilot vehicle navigation tasks.

The HPM validation methodology is being developed to provide a framework and scaffolding to promote calibration and validation of human performance models along this continuum. We are developing this framework by working systematically from two given points of reference--the model specification and the empirical performance situation--in order to formulate the integrating representational framework that provides a reliable mapping between the endpoints. At the empirical endpoint, the methodology focuses on a taxonomy of observable actions which permit automated data collection (e.g., keystroke actions, voice utterances, body movements) of data that are conceptually relevant to the model evaluation process. At the modeling endpoint, the methodology identifies events that are 'mappable' to the observable actions and also to evaluation criteria.

While many validation studies have been conducted to calibrate and validate human-performance models against relevant empirical data, the complexity of the many factors and variables involved makes it very difficult to develop general interpretations of the results. One major challenge is presented by the kind of 'bundling' that typically occurs in the development of a human performance modeling application. Typically, the same organization and people who designed and produced the modeling tool/framework are also responsible for the engineering application of the test case, thus making it difficult to attribute any observed results to the modeling tool/framework as opposed to the engineering skills of the project team in accomplishing the immediate application. Also, this same team is sometimes responsible for collection of the empirical data to be used for model evaluation as well as for conducting the statistical analysis and evaluation. Alternatives to this situation are often difficult to arrange because the complexity of the models makes it costly for people not familiar with the models to conduct these types of analysis.

5.3 Concepts of Calibration vs Validation

The issue of calibration vs. validation is concerned with a distinction that may or may not be judged to be necessary. Indeed, the landmark review of human performance models sponsored by the National Research Council (Reference 1) treats calibration and validation as synonymous, with 'calibration' referring to a type of validation. However, we would like to point to a distinction between our preferred usage of these two terms which we believe also serves to clarify this problem area. Our view is that calibration should refer to the process by which a model is adjusted in order to best fit one or more empirical data points. Validation, on the other hand, should refer to the more general process by which we use the comparison of the model to one or more empirical data points in order to speculate about the likely comparisons that would obtain between the model and empirical data across a broad (but bounded) space of possible conditions. In fact, the explicit calculations and comparisons that are made may be essentially the same between the calibration and validation processes, but plans and conclusions can be quite different. Whereas we should be able to calibrate a model to any collection of empirical data points where the model can be applied, those data points may not be very suitable for validation purposes because they may not provide a very good sampling of points over the space for which we wish to apply our judgment of model validity. The ideal procedure for validation is to select a set of points in the space of possible simulation and empirical conditions such that confirmation of good or acceptable model predictions at those points will justify extrapolation of a similar goodness of fit throughout the remainder of the space. But these selected points also need to be segregated into 'calibration' points and 'evaluation' points, at least during the process of the validation. At the calibration points, the empirical data is used to adjust the model in order to obtain the best possible fit. The parametric conditions at the evaluation points are then used to generate model predictions without knowledge of the empirical performance data corresponding to those points. The comparisons of the model and empirical data at the evaluation points then serve to provide the estimates or measures of model quality across the total space of conditions of interest. Of course, after the validation comparison is made, the evaluation points can be turned into calibration points by further adjusting model parameters to achieve best fit at those points as well.

Thus, it appears that the main difference between the processes of calibration and validation lie in the sampling strategies. For the purpose of calibration, we assume that you would generally pick your calibration points as close as possible to the points where you intend to use the model for predictive decision making purposes. For validation, the goal will generally be to establish calibration and test points (which subsequently become calibration points) in some sort of uniform distribution across the application space of interest. Our presumption here is that validation is a general process that is intended to support a broad range of future applications of a model, whereas calibration is instead focused on just one or only a small number of such model applications.

As a step toward formalization of the calibration and validation concepts, consider the following nomenclature:

Let Pc be the variable that defines personnel characteristics (a vector quantity).

Let Ss be system specifications for own system.

Let Cs represent context specifications (including other systems and the environment).
Let Sc represent the scenario specifications.

Then we can define the empirical performance data that is generated by observing this person in this system and environment operating in this scenario as $\mathcal{E}$(Pc, Ss, Cs, Sc). Then additionally, we let Sm represent our host simulation specs and let Cm represent our cognitive model specs. Then we define the performance data generated by the simulation (Sm and Cm combined) with the indicated conditions as $\mathcal{S}$(Sm, Cm, Pc, Ss, Cs, Sc). Then we can further define the measure function $\mathcal{M}(\mathcal{E}, \mathcal{S})$ as the measure of the goodness of fit between $\mathcal{E}$ and $\mathcal{S}$. We can now define calibration as the process of adjusting CM specs, model parameters, etc. in order to optimize $\mathcal{M}$ for a given context or (discrete) range. And validation can be defined as our 'speculation' about the limits or distribution of $\mathcal{M}$ across a completely specified range of contexts.

Based on all of this background, we can now offer several speculative conjectures:

(1) We should interpret the difference between calibration and validation of CMs to be representable as a continuous range of variation along a dimension from pure calibration at one end to pure validation at the other extreme, with intermediate points corresponding to intermediate degrees of extrapolation or interpolation from the established empirical data points.
(2) CM frameworks can't be validated (or calibrated). They are really just software environments, and may not even incorporate specific micro-models.
(3) CM frameworks influence efficiency and transportability of model implementation, but not validity.
(4) Both calibration and validation are specific to the specific output data generated by $\mathcal{E}$ and $\mathcal{S}$ (the performance criterion data) and the specific measure functions used for comparison. Thus, a model that has been validated for prediction of performance times or workloads, in some fashion, is not necessarily thereby valid for predictions of performance accuracy or decision quality.
(5) Like the CM framework, the identity and capabilities of the user/analyst who builds the model should be irrelevant to the validation, but, like the CM framework, an influence on efficiency of model development.

This idealized view of calibration and validation becomes significantly complicated by the consideration of the economics of data creation. It is generally very costly to obtain data both for empirical data points and for the corresponding model data points. Sometimes it is possible to establish multiple data points through simple parametric variation of easily manipulable aspects of the experimental environments and the corresponding models, but many of the more important variations will require major changes to system designs, environmental context, and human performance strategies. On the empirical side, major changes in design and context will also require new training of experimental subjects and possible recruitment and training of totally new subjects, in addition to possibly major changes to the experimental testbed facility or transition to a different facility. On the modeling side, major changes to system (especially user interface) may require fundamental reassessment of human operator strategy and tactics entailing a costly model revision effort going back to new CTA studies with SMEs. These considerable costs and associated programmatic delays mean that very few data collection efforts are ever pursued to even try to collect comparable empirical and CM model data points,

and where such efforts are pursued they are done so on limited R&D efforts (such as on the AMBR project which is the most relevant and complete effort in this vein) so very few data points are collected. Of course, one of the main reasons to use the model in support of SBA decisions is to avoid the need for empirical design implementation and performance data collection, so very few empirical data points are established. However, in an ideal application of CMs and associated simulations in support of SBA, there would be a relatively large number of CM/simulation data points generated in order to explore the full space of design and application possibilities as dictated by acquisition goals and requirements.

Costs vs benefits are, in fact, a central consideration throughout any acquisition process, and decisions about the use of simulation and CMs must be understood relative to this trade-off. The SBA managers key decision, from our perspective, is not whether or not to use a simulation+CM approach to evaluate design options at each design decision point, but rather what approach (es) to use from the full array of available approaches. Expanding on our prior breakdown, the plausible alternative approaches for design decision advice seem to include at least the following:

(1) Subject matter expertise from SMEs with either engineering or operational expertise (or both), unguided expect for presentation with the core question;
(2) Subject matter expertise from SMEs, with coarse methodological guidance for elicitation and processing of component estimates at the level of major design components – available methods include the Navy Hardman method (Reference 54) for structuring of anchoring-and-adjusting estimation of design feature impacts on performance or the more recent Navy performance metrics methodology (Reference 55);
(3) Subject matter expertise from SMEs, with guidance at the level of a cognitive task analysis and workload estimation using standard workload estimation scales (Reference 60);
(4) Subject matter expertise from SMEs based on experience in executing simulated scenarios with the new design in a virtual environment simulation of the system design and environment (where the simulation could consist of various combinations of real 'hardware' and virtual implementations of relevant components);
(5) Empirical performance data from the performance of SMEs in virtual implementations;
(6) Model predictions based on performance of CMs in the simulation environment (presumably the same core simulation that would be used in the constructive virtual implementations with real SMEs as human operators).

These options are not really distinct alternatives, but more like optional stages of analytic sophistication, with each succeeding subsuming most of the resources and costs of the preceding stage, but with some significant exceptions. Basically, some SMEs are required for all approaches, some analysis methodology is required for all except the first stage, and some core simulation capability is required for stages 4, 5, and 6. So the problem for the SBA manager is to determine, relative to the array of design decisions, what the costs and benefits are for each of the stages of analysis options. On the 'benefits' side, the manager needs to know about the likely quality of the design advice that is provided (i.e., a measure of information 'validity'), while the 'costs' side requires consideration of all of the expenditure, investment, and schedule factors. It is important to recognize that the costs for development and maintenance of simulations and CMs must be understood to warrant amortization over multiple decision points in the acquisition program that will be supported, and possibly extending to other related acquisitions.

5.4 Aspects of HPM and CM Validity

Human performance can be described in many different aspects and levels of detail according to the behavior representation tool or method being employed. Task network representations can be as simple as identifying just the start and stop times of all of the discrete tasks that make up behavior, but they can also provide a variety of amplifying data and structure, such as workloads and outcomes associated with tasks. Cognitive modeling architectures will typically provide full descriptions of task dynamics along with detailed descriptions of the behavior of component cognitive and manual processes, such as visual perceptions and eye movements. It would be appropriate to construct a sort of "common denominator" type of behavioral description language to encompass all of these modeling techniques, such as Ianni (Reference 61) and Badler et al. (Reference 62) have offered for the slightly different context of behavioral representation for graphic human models. This type of common language is worth considering for development to assess unique HPMs within this general validation framework. Further, the use of such a common language to describe model elements could facilitate the establishment of links between T&E HPM requirements and available HPM options. It could additionally create potential for generalizing empirical results across HPMs and HPM elements of the same type, thereby enhancing the efficiency with which HPMs may be validated.

The typical situation with HPM models is that the SBA design engineer wishes to obtain predictions or evaluations regarding human performance in a new system design and recognizes that some previously built model is somewhat relevant, but not precisely applicable without some adaptations. Hopefully, the legacy model will come with some sort of validation data in terms of comparisons of model predictions to the empirical performance for the target system (though we should also recognize that such empirical validation data is typically far from ideal if available at all). It is necessary to examine the kinds and magnitudes of the adaptations that need to be made to the legacy model in order to judge the degree to which any prior validation of the model should be carried into the adaptation.

Corresponding to the many aspects of the complexity of cognitive models and system/environmental simulations, there are many types of model adaptation that pose different issues for the considerations of model validity. First, there are a number of areas where changes can be made that are actually external to the cognitive model, but nevertheless have important implications for model validity:

- System design features – Features of the system design can have a wide range of impacts on the cognitive model, from no impact to major changes when the altered features involve a complete revision of the user interface. In cases where a new user interface incorporates a radical new technology (such as natural language interaction) for which the cognitive model has not previously been applied, it may be necessary to recognize major questions about the validity of the model in the new application context even if validity was firmly established in the baseline system context.
- Environmental and other performance moderating features – Many very different environmental factors and components can be important in the implementation and validation of a cognitive model. On one side there are all of the diffuse external factors such as weather, vibration, visibility, etc. Issues of stress, workload, fatigue, sleep cycle,

drug effects, etc. can also serve to influence performance as a consequence of external factors, though these kinds of factors can also be interpreted as internal organism or model characteristics. And under this broad category, we must also consider the roles of other players and systems in the simulated world, including teammates, adversaries, and other participants in missions and scenarios of interest.

- Scenario features – In addition to system and environmental features, the characteristics of the simulation scenario can result in major changes to the aspects of a cognitive model that are invoked and exercised in a simulation. So a model that is observed to conform to ground-truth in one scenario may deviate considerably in another scenario. While it can be difficult to anticipate all relevant scenario features and how they will influence model validity and performance, it is certainly possible to analyze scenario composition and develop techniques for scenario generation and selection so that model validation can be conducted with strategically chosen scenario conditions in order to maximize the generalizability of the validation results.

Then there are many more areas within the cognitive model that have strong implications for the validity of the adapted model:

- Knowledge base content (declarative and procedural) – Cognitive models employ extensive knowledge representations, including multiple different types of knowledge according to the type of ontology on which the model is based. Cognitive models will typically incorporate some representation of objects in the environment (including own system and other friendly and non-friendly systems), declarative knowledge about those objects, goals to be considered for performance by the modeled human, and tactics, techniques, and procedures (TTPs) for accomplishing those goals. Changes to any of these elements of the knowledge base can clearly influence the validity of the model, though case-specific analysis of the situation can potentially serve to estimate the likely impact.
- Micro-model structural changes – Cognitive models typically include some form of micro-models to describe component aspects of performance such as perception, motor action, decision making, etc. In many cases, these micro-models can be selected from the research literature where substantial empirical evidence is provided concerning the general validity of the micro-model. The most commonly used micro-model for hand and foot movement ("Fitts' Law") is a prominent example (Reference 63). In fact, the case of Fitts' Law offers an interesting example of a kind of micro-model change that might be encountered. Fitts' Law describes how body movement time (typically hand or foot, but also translating to computer pointing devices such as mouse or trackball;(see References 64, 65, 66 for several key extensions and elaborations) varies as a function of movement amplitude (i.e., distance or extent) and required accuracy of movement (such as the size of the terminal target that the hand or foot or cursor has to stop at). But if the modeler were to need to have the model predict the distribution of movement errors in an area where Fitts' Law was initially being used as the controlling micro-model, then it would be necessary to switch to a different kind of micro-model for that category of movement that would treat the terminal movement error as an output rather than as an input to the model (e.g., see References 67, 68). If that were the only change to the simulation, then a key question for validation would be how to combine the separate validation evidence for the prior simulation with Fitts's Law micro-model and the distinct

empirical laboratory research evidence for the validity of the new body movement micro-model.

- Micro-model parameter changes – Changes to micro-model parameter values are usually interpreted more in the context of calibration of models and simulations rather than with regard to validation of those models and simulations. By systematic or trial-and-error adjustment of parameter values, we incrementally refine the fit between model predictions and empirical reference data. But it is important to recognize that parameters may be adjusted predictively outside of the calibration context in order to make the simulation apply to a key alteration in context (e.g., postulating that the multiplicative scale factor in a Fitts' Law micro-model would vary in proportion to the fatigue level of the human operator). Of course, micro-models can be discrete as well as continuous in their range or effect, so some micro-model parameters can signal whether or not whole micro-models or component feature dynamics (e.g., learning, fatigue, stress, emotion, etc.) are to be invoked. So a micro-model parameter change could represent a move into new model prediction territory that warrants a new consideration of model validity, or it could be a small move within territory that is already well mapped with regard to validity.
- Priorities for goals and tasks – Goal and task priorities represent a special and very important type of micro-model parameter since they are representative of the attention model that is employed by the cognitive model. Whether or not it is explicitly articulated in the documentation of a cognitive model, every cognitive model must have some sort of attention model (or micro-model) to determine how to distribute processing 'attention' in the course of simulation execution. Many CMs assume that attention is an 'all-or-none' process where only one task or procedure can be active at one time but that attention can be switched back and forth between multiple tasks with some frequency. In this case, some priority parameters are needed to drive the decisions as to when to interrupt for switching and which task to switch to at such choice points. Alternatively, other models will postulate the option for multiple tasks to be executing at the same time, but with constraints on total capacity (how many tasks) and possibly individual task performance characteristics being governed by other types of attention model parameters (e.g., percent of attention capacity, or capacity by separate performance 'channels', required for performance of the task). Changing attention model parameters can be done in the course of model calibration, but it can also be done, as with other micro-model parameters, in a more extrapolative manner that would suggest the need for additional validation.

Decision makers in an SBA program ideally should want to select HPMs and CMs to incorporate into the SBA program as part of the initial planning process. Key to HPM and CM selections by SBA decision makers would be information about the validity of available HPMs and CMs. Subsequent decision points at which HPM and CM validity would be pertinent would be prior to the simulation-based evaluation of each iterative version of the system under development. Specifically, decision makers may need to consider the incorporation of additional HPMs and CMs or the validity of previously adopted HPMs and CMs in light of changes in the simulation environment, the system to be evaluated, or system evaluation objectives.

As noted previously in this report, the issue of model validity is complicated by the number of factors that can come into play. Overall model validity may be impacted by myriad factors including those listed below:

- *Computerized model validity, also designated as verification* (e.g., Reference 69) –
  - the extent to which the architecture or tool used to develop a model permits the accurate representation of various aspects of human performance.
  - the extent to which the conceptual model has been implemented correctly.
- *Conceptual model validity (e.g.,* Reference 69) *or construct validity (e.g.,* Reference 19) -
  - the extent to which the architecture or tool used to develop a model is consistent with theory about human performance and cognition.
  - the extent to which the model is consistent with relevant theory about human performance and cognition.
- *Data validity* (e.g., References 19, 69) - the quality of human performance data from which a model is derived.
- *Face validity* – whether a model's performance is perceived to be realistic by human simulation participants or others who are knowledgeable about the performance domain (e.g., Reference 69).
- *Predictive validity* (e.g., Reference old-34) - the extent to which model performance is consistent with and predictive of human performance under conditions for which data have not yet been fed into the model.
- *Operational validity* (e.g., Reference 69) - the extent to which the model produces outputs that meet the host simulation system's needs in terms of output type and degree of accuracy.

As an example starting point in the process of selecting HPMs and CMs for use in an SBA program, SBA decision makers might first want to know what HPMs and CMs are available that simulate the categories of performance (e.g., decision making, perception, hand movement, marching, etc.) or moderating effects on performance (e.g., fatigue effects) that are needed for the acquisition program. Thus, if SBA decision makers want to assess how well the system to be developed disrupts an enemy integrated air defense system (IADS), they might want HPMs and CMs that simulate track processing and decision making. If they want to assess whether a system improves coordination in a strike mission team, the decision makers might want HPMs and CMs capable of demonstrating perceptual, information processing, attentional, and working memory limitations; and demonstrating the effects of those limitations on factors such as information sharing, anticipatory support among team members, and situational awareness of team member positions and activities.

To assist SBA decision makers with the identification of relevant models, HPMs and CMs should be indexed or otherwise presented in a manner that facilitates comparison and assessment. Efforts in this direction include Silverman's Human Behavior Model Anthology (HBMA; (e.g., Reference 70). Silverman and his colleagues have developed a catalog of HPMs and CMs indexed according to measurable descriptors and information about their derivation, their range of utility, and implementation lessons and suggestions. The HPM/CM Validity Assessment Scale shown in Table 1 was developed by Silverman and his colleagues as a means to characterize the validity of model construction within the HBMA (and may be used to assess a model's data

validity). In a separate relevant effort, Harmon (Reference 71) developed a taxonomy of HPM/CM capabilities to support Human Behavior Model requirements specification and that also may be used to index HPMs and CMs in support of SBA model selection (Figure 4). Harmon's taxonomy defines three categories of model characteristics: Non-Cognitive Factors, Cognitive Capabilities, and Military-Specific Functions. Non-Cognitive Factors include physical factors such as weapons and environmental effects and psychological factors such as aptitudes and emotions. Cognitive Capabilities consists of three primary categories—situation understanding, plan construction and plan execution (see Figure 5 for a break-down of the Cognitive Capabilities category). The Military-Specific Functions refer to specific performance tasks across the entire spectrum of military operations.

Table 1. Scale for assessing the validity of studies conducted to derive HPMs and CMs from Reference 70.

Scale	Degree of Value of Literature Item for Constructing HPM/CM
5= Very High	Model provided with backup data sets.
4= High	Could develop models directly from the data in this study.
3=Medium	Some preliminary data for initial model construction; more data needed.
2= Low	Theoretical model suggested from which an ungrounded model could be derived.
1=Very Low	No valid data in this report for PMF construction.
0=None	Irrelevant to the model construction process.

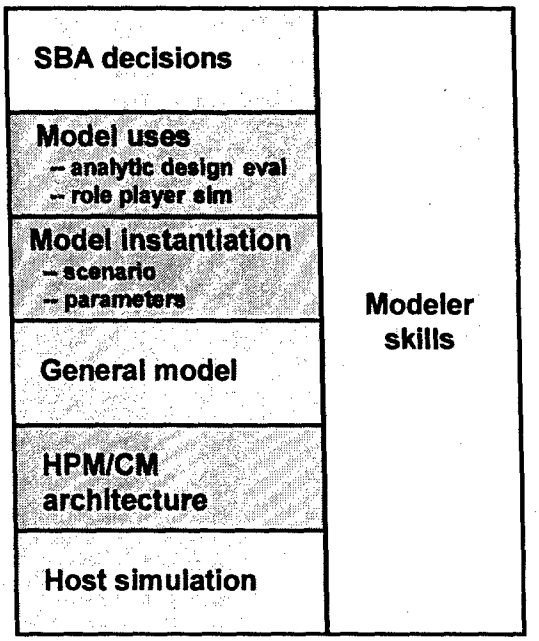

Figure 4. Layers of Model Generation & Specification

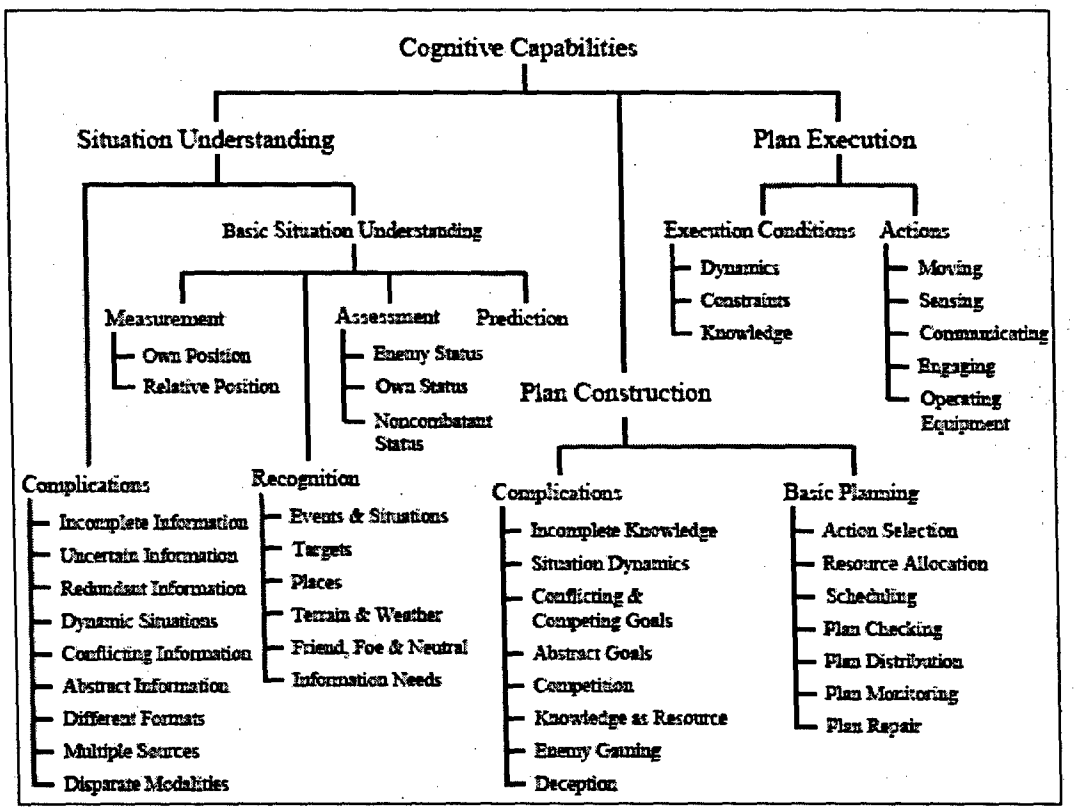

Figure 5. Cognitive Capabilities Specified by the Human Behavior Model (HBM) Requirements Taxonomy Proposed by Harmon (Reference 71).

Once a set of candidate HPMs and CMs has been identified—in this case, HPMs and CMs that simulate the desired types of performance or produce the desired moderating effects on performance—SBA decision makers will want to further evaluate and narrow down the choices with respect to a number of other practical issues. SBA decision makers may consider computerized model, conceptual model, and data validity at this point. However, they are more likely to be interested in the face, operational, and predictive validity of the candidate models. Face validity is likely to be an important factor to SBA decision makers because it is such a salient factor and can significantly impact trust placed in the model and its effects by simulation participants and stakeholders. Operational validity is an especially critical factor for two reasons. First, one of the factors to which operational validity refers is the type and form of HPM/CM outputs. These outputs need to conform with the types of inputs the simulation system can accept and with the types of measures that are needed for the purpose of evaluating the system being developed. For example, if a model produces output in the form of a throughput measure that combines accuracy and response times, it would not facilitate the simulation-based assessment of decision accuracy or target detection times associated with a C2 process that is under evaluation. Second, operational validity refers to the fidelity and accuracy of a model in terms of whether it meets the fidelity and accuracy requirements of the simulation system or acquisition program. Thus, a wingman CM may not require a high level of fidelity or accuracy if that model exists only to accept and blindly follow commands from a pilot CM or HITL who is directly interacting with the system under evaluation. Finally, predictive validity becomes an important consideration if a CM or HPM with a high level of accuracy is required—for example, if key decision evaluation measures include CM performance characteristics and outputs.

In addition to the categories of validity listed above, SBA decision makers might consider other factors that can have major implications for HPM and CM validity within a given SBA simulation environment. These factors include:

- whether and how well a model is able to interface with the simulation environment;
- the extent to which a model is able to function as intended within a given simulation environment (something that may be affected by, for example, the ability of the simulation environment to provide the model with necessary inputs);
- the scope of model validity within different categories of context (e.g., domain, scenario, environment, and evaluation context);
- the ease with which an HPM or CM can be adapted in order to meet a given SBA requirement (e.g., what type of expertise is required, was the model built using a modeling tool with an intuitive user interface, will a single change require a chain of modifications, etc.); and
- the impact of model adaptations on overall validity.

The latter three factors above are especially pertinent to the initial selection of CMs and HPMs for an SBA program if a model will be used in a context that differs somewhat from contexts in which it has been validated or if it will be adapted to better meet the requirements of the SBA program. Similarly, these factors are pertinent to the re-use of CMs and HPMs across design spirals and evaluations within an SBA program. In particular, it is likely that different aspects of the simulation system and evaluation context will change across program spirals and those changes to HPMs and CMs would be necessary as a result.

The implications for validity introduced by such changes can be complex and difficult to assess. In Table 2, we present four context factors and describe their implications for model validity. Changes made to a CM may include changes to its perceptual/input capabilities, output capabilities, knowledge-base content, parameter values, attentional and prioritization scheme, or procedural content. Such changes may be necessary in response to certain context changes and would almost always call for the re-assessment of model validation with respect to all validation categories listed above. In cases in which significant model changes may be required, it may be reasonable to consider the adoption of alternative models.

Ideally, HPMs and CMs chosen for use by an SBA program will be maintained within the program's collaborative environment (CE) and will be evolved in concert with all other technologies supporting the program in response to various changes and across time. Model maintenance within the CE would need to accommodate the maintenance of model validity, particularly in response to changes to the simulation environment, system or process being developed, and CM/HPM.

Table 2. Context factors that may impact previously established validity of a CM or HPM.

Context Factor	Implications for Previously Established Model Validity
Domain Context	This factor refers to the types of work or operational domains for which a model has been validated and the scope of that validation. With respect to domain-context scope, a model might be valid for a narrow context as in the case of a model of an F-16 pilot flying air combat missions in a two-ship configuration versus for a broad context such as a track processing model that adapts across rules of engagement, operation, and C2 organization (e.g., across Battle Control Center [BCC], AWACS, and E-2C).
Scenario Context	This factor refers to a model's robustness across simulation scenarios. Has a model been validated using only a subset of scenarios within a given domain and simulation environment and if so, does model performance in those scenarios generalize to model performance and validity in other scenarios? The characteristics of a simulation scenario can result in major changes with respect to the aspects of a cognitive model that are invoked and exercised in a simulation. Consequently, a model that is observed to conform to ground-truth in one scenario may deviate considerably in another scenario.
Environment Context	In certain cases it may be necessary for a CM or HPM to be responsive to the environmental context within the simulation environment and to be validated with respect to this responsiveness. For instance, a CM or HPM may need to take factors such as weather, vibration, visibility, time of day, and the numbers and types of simulation entities into account. When this type of responsiveness is required and an unexpected change in the environment is introduced, model validity may need to be assessed with respect to any changed environmental factors and model adaptations may be required.

Evaluation Context	The features of the system or process under development in an SBA program can have a wide range of impacts on a CM used to support evaluation. For instance, in the case of a major change to a system design, the nature of inputs from the system to the model and vice versa during a simulation-based evaluation might change dramatically. In order to support the evaluation of the design, the model would need to be adapted and its operational validity re-assessed. Similarly, if a radical new technology such as natural language interaction were added to a system's user interface design, it may be necessary to consider major questions about the validity of the HPMs and CMs used to evaluate the system, even if validity was firmly established in the baseline system context.

6. CM Validation in the Emerging SBA Context

6.1 Anticipated Relationship of CM to Virtual Environment

There is tendency to think of the validation of a model as a discrete event involving the comparison of model outputs/predictions with corresponding empirical data. For the case of CMs, however, there are several compelling reasons to consider CM validation as a complex evolutionary process. A primary source of problems is the high relative cost of developing most CM applications and the substantial additional costs that are required for empirical data collection, especially if the data collection process must be supported completely by the validation activity. Along with the costs involved, there are also problems deriving from the continual incompleteness or instability of most CMs – there are always important new features and capabilities that are "in development." Because of the layered composition and the diversity of usage requirements for CMs, it seems unlikely that very many CMs will be 'frozen' in some packaged form and validated through comparison with empirical data and offered for re-use. While such focused validation is possible, it is expected that efforts to assess CM validity will be concerned primarily with determinations of the likely validity of the model when it is somewhat reconfigured and translated from the specific context in which the validation was conducted. One application where validation of fixed models is most likely to occur is the area of synthetic agent players for virtual environment experimentation. In these cases, the VE-HITL facility will almost necessarily precede any CM development, so empirical data will be available for calibration of the model, and peripheral role players can be expected to be usable in fixed, generic form across many application environments and purposes.

Figure 6 illustrates the view that we have developed of the various elements involved in CM performance generation and hence in considerations of validation. A host simulation is provided for the representation of all of the non-human aspects of the systems and environments of concern – own systems/platforms, other friendly force (FFOR) systems/platforms, opposing force (OPFOR) systems/platforms, and environmental characteristics and procedures (e.g., terrain, weather, noncombatant elements, etc.). The host simulation will also typically provide a specification for all key scenario events in order to drive and coordinate the simulation in order to address application purposes. Although the host simulation will likely be constructed using a variety of software tools, it will still all be coordinated and managed through a common simulation architecture such as DMSO's High Level Architecture (HLA).

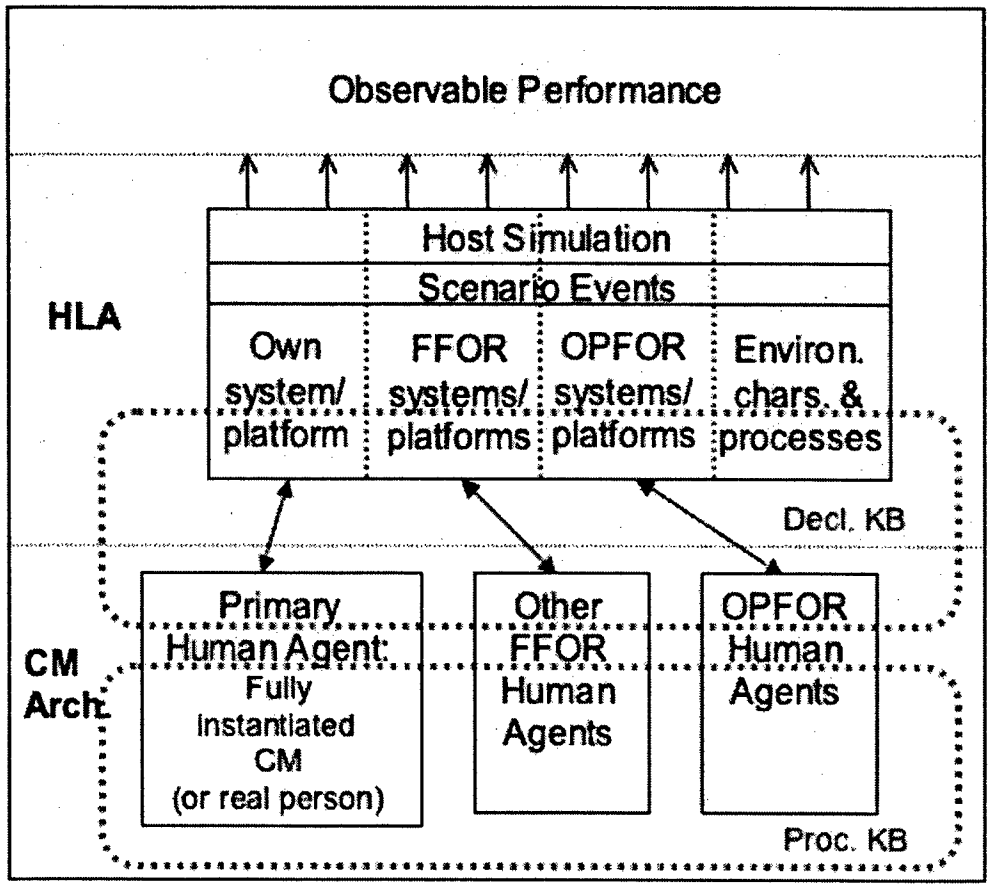

Figure 6. Architectural View of Host and CM Simulations

The CMs are shown in Figure 6 as multiple processes that connect separately to the host simulation and are all implemented in a common CM architecture. In fact, there is no reason that the multiple CMs couldn't each be represented in a different architecture, though we assume that the most common situation would be for all of the CM agents to employ the same CM architecture. It is important to note that the agents that we are considering in this context for possible validation must all be fully instantiated specifications that are capable of effective execution in coordination with the host simulation. In particular, there is typically a very large amount of agent specification information that is not inherent in the underlying CM architecture. Some of the CM agent specifications will constitute portions of the "declarative knowledge base" characterizing the relevant context-specific factual information that is needed to perform the job and tasks of the agent. To the degree that this declarative knowledge overlaps with the components and processes of the host simulation, this CM knowledge base may be shared or linked with the host simulation, though there may also be declarative knowledge that the CM agent generates and uses without being represented in the host simulation (such as self-knowledge about the status of own tasks and performance resources). Procedural knowledge must be provided for all CMs in order to specify context-specific logic for behavior, identifying the goals and methods that are invoked to perform tasks. But, unlike declarative knowledge, procedural knowledge will be contained only in the CMs and not shared with the host simulation. The figure also indicates that the observable performance that is of interest for CM validation purposes is generated primarily from the host simulation rather than as the direct output of the

CMs. That is, we expect to be primarily interested in what the agent does with and through the system/platform as opposed to any 'local' CM cognitive activity. Of course, we are occasionally concerned with cognitive measures for constructs such as workload or situation awareness that are not otherwise represented in the host simulation, so such cases will require direct generation of CM outputs that can be compared with corresponding construct estimations generated by real people.

6.2 CMs and Application Validity

Certainly part of the application specificity that is entailed in the concept of "application validity" is represented in all of the domain and task specifications that we illustrated and discussed with Figure 6. However, there is another important aspect of the application that is also needed in order to pursue the validation of the CM – that is the understanding of the purposes and goals of the user or SBA manager who needs to judge the validity of CMs and make decisions about their uses for SBA purposes. In general, we expect that the question of validity will focus on the same aspects of performance that are of primary concern for the SBA issues. For example, in one situation the SBA design issue might be concerned with the ability of typical system operators using a current interface design to perform the required tasks within an acceptable time period or error rate. So we would want to validate the model with respect to performance times or error rates accordingly. In other cases, we might be primarily concerned with issues of usability, workload, situation awareness, manning, etc. And it is also noteworthy that the specific criteria for task success, task completion, and errors will be uniquely defined by the system/platform performance requirements.

While we expect that the SBA program that is using and validating the CM will focus on the performance criteria that are of primary interest for the immediate program application, there are at least two reasons to simultaneously strive for the broadest possible program of data collection and validation. First, it should be recognized that no one will know all of the design evaluation issues that may arise at later stages of the acquisition program. While initial interest may be just in gross workload and manning estimations, later stage concerns may re-focus on the effects of individual differences in the user population on system performance. Secondly, it should be recognized that more general data collection and validation of a CM in one program may incur relatively minor additional costs to that program while providing valuable benefits to future programs. Thus, planning for CM validation in SBA presents a difficult and potentially consequential trade-off that warrants detailed attention and structured guidance.

In discussing SBA applications of CMs, we tend to think mainly of using the CMs as diagnostic tools that we use to generate surrogate human performance and thereby directly evaluate the performance consequences of design features. However, there is another important use of CMs in SBA, and in fact an application that is probably more active and mature than the diagnostic evaluation role – this is the use of CMs as synthetic agents (synthers) in virtual environments in order to provide sufficiently realistic behaviors for peripheral role players and to enable the conduct of experimental evaluations of design features with real people in the primary agent roles. We will refer to this type of experimentation, with or without synthers, as virtual environment human-in-the-loop (VE-HITL) experiments. Applications of CM synthers for VE-HITL experiments will probably appear very similar to uses of synthers for simulation-based

training applications (e.g., References 72, 73), though the desired performance characteristics and validation criteria are likely to be somewhat different, though not necessarily very much so.

Clearly, the VE-HITL experiment is a natural standard for comparison against the 'constructive' CM application of Figure 6 in order to assess CM validity. However, the VE-HITL standard may be conceived in a variety of ways, especially when multiple human roles are involved. The VE-HITL concept entails at least one real human playing the primary system operator/user role while all of the peripheral roles could be represented by either CMs or other real people in various combinations. At the same time, the VE side of the concept also affords a range of alternatives in terms of how all of the system and environmental components are implemented. Indeed, a distributed architecture like HLA is commonly used to support configuration of a distributed 'host' simulation environment, with different systems, platforms, and environmental components being simulated at different sites with different software tools. In any case, it seems clear that the ideal referent for all validation purposes will be maximally similar to the constructive CM simulation with the sole exception of the CM replacing the real human in the role of primary interest. If we vary other components between the CM simulation and HITL referent, then we are likely to encounter difficulties in determining which component differences are responsible for any differences in performance between the two cases. In this vein, the use of CM synthers rather than real people in the peripheral human roles is preferable because it allows for better control of the behavior of the peripheral role players across the conditions to be compared. However, if the goal is to evaluate/validate how well a CM interacts with real people in various peripheral roles, then it may be more appropriate to place real people in such roles in both of the conditions being compared.

6.3 Validity as Correspondence

Here we address the concept of model validation in the sense of correspondence with empirical reality rather than the sense of theoretical legitimacy. We acknowledge that we can develop some appropriate confidence in a model by verifying that it conforms with a theoretical understanding of the organization of behavioral processes that we have identified through various empirically grounded research. However, this path to validation seems essentially to be based on the idea that we can validate one model by establishing its accurate correspondence with one or more other models that have been separately validated. We choose instead to focus on the explicit correspondence of model-generated performance with empirical performance as the primary basis of model validation.

Validation of CMs, just as for validation of other types of simulation, must start with a recognition of the distinction between verification of the software implementation of the model/simulation and validation of the specification for the model/simulation concept. The distinction and prerequisite relationship between verification and validation of simulation models has been a prominent assumption in the simulation community (e.g., References 1, 74). Essentially, the term 'verification' is used to denote the confirmation that the software is an accurate representation of the conceptual specification of the model or simulation. And the term 'validation' pertains to the correspondence between the performances of the software implementation of the model/simulation and the associated empirical reality, so validation may be considered to presume or subsume verification. In general, we will want to establish the

verification of the simulation software before we attempt to pursue its validation. In the simulation architecture view of Figure 6, we view the verified software implementation of the simulations as an essential layer under all else.

A difficulty seems to arise when we attempt to identify what it is within the structure and content of the CM that we are validating when we establish the quality of the correspondence between simulation performance and empirical performance. To what degree are we validating the general CM architecture that is used in the simulation, or knowledge base, or the component human performance models? Or is it just the specific simulation for the current domain in all of its details that is validated? Surely we are striving for a fair degree of generalizability in seeking validation of a model. We want to know that a new application of the model in a new domain will produce a reasonably accurate prediction of real-world performance.

6.4 Calibration and Validation of CMs

It is useful to define the process of 'calibration' of models as distinct from that of validation. In general, calibration suggests that the parameters or specifications of the model are adjusted in order to optimize the fit of the model to empirical data that is provided to the modeler. Validation, on the other hand, refers to how well the model results are likely to fit empirical performance prior to any such adjustments. Thus, for validation purposes, we are primarily concerned just with the initial performance of the model at the calibration points prior to any adjustments of parameters or specifications. This condition presents some difficulty in most real model development processes because the modelers typically have some estimated or partial performance data to which to fit their models in most situations, so the model is typically fit to that partial or estimated result in a first iteration, and then subsequently adjusted to fit the complete empirical performance results when they are available. This is what occurred in the recent 'run-off' between several alternative CM frameworks and teams in the Air Force Agent-based Modeling and Behavioral Representations (AMBR) program. AMBR engaged four modeling teams (representing the CM frameworks of AMBR, DCOG, EASE, an ACT-R/Soar/EPIC hybrid, and iGEN®) to develop models for human performance in a game-like task loosely based on air-traffic control operations (References 75, 76). The objective of each team was to develop and configure a model, using their CM framework, to predict the performance of a group of test subjects. In fact, the modeling teams were provided performance data for a partial group of test subjects at the outset, and even if that was not provided, each group could easily have generated similar data because all groups had copies of the game testbed. Although, data about the calibration fit of each model to the partial data was reported for AMBR, there were no reports of model predictions prior to any adjustments. And for good reason, because models are typically developed and tested incrementally against whatever data or estimates are available.

It is additionally noteworthy that AMBR presented the modeling teams with three distinct modeling challenges that serve to illuminate our concerns about validation and calibration. In the first challenge (Experiment 1), the teams were asked to construct models to predict the multi-tasking response times of game players across workload and aiding conditions and also to predict the workload estimates of the players. In the second challenge (Experiment 2, Part A), the teams were asked to extend their models to predict game-player performance (time and errors) in a

category learning elaboration of the basic tasks. In the third challenge (Experiment 2, Part B), the models were run in a new condition (extrapolating the category definitions from Part A) without allowing the teams to make adjustments. In terms of the distinctions that we posed above, the first and second challenges served just to provide calibration points and the third challenge seemed to provide a real test for validation. In fact, the general consensus appeared to be that all of the AMBR teams failed in the third challenge (References 77, 78), but there are plausible reasons to consider the third challenge as a rather extreme extrapolation from the prior cases and, thus, not necessarily a very significant indication of model validity.

Thus, it would seem that our interest in validation can be viewed in terms of extrapolation (and/or interpolation) of model predictions across the space of domain variation of interest. Thus, we want to know how reliable the model results will be when we make changes to the user interface, the task requirements, the protocols for training and learning, etc. There are many dimensions in which the domain specification can vary, but for illustrative purposes, we will collapse all of these into two dimensions in Figure 2 (the two dimensions in the horizontal plane) in order to illustrate the issues of concern. For this example, we will present some of the AMBR data generated by the iGEN® team (Reference 79), but with the assumption (contrary to our argument above) that Challenges 1 and 2 provided true validation data in the comparison of model and empirical data. In general, we contend that the job of model validation is to develop a sort of "response surface" prediction of "model accuracy" across the multi-dimensional design/application space, based on the established pre-calibration data points that are provided. The contrasting bar charts that are stacked on top of each of the Challenge points in the design space each represent the comparisons of model versus real people for key criteria for that challenge (i.e., response times for Challenge 1; concept mapping categories for Challenge 2; treatment of new category exemplars in Challenge 3). In a more general sense, our contention is that the performance comparison at each of these "validation design points" should be considered as a vector quantity of all of the component comparisons for the relevant criteria.

Further developing on the view of Figure 7, we envision that validation of CMs should occur through the development of data at substantial numbers of validation data points (corresponding to the 3 challenge points in Figure 7) so that extrapolations of prediction accuracy at new points in design space should be reliable. Here it is important to remember that there are typically a much larger number of dimensions in the design space than the two illustrated in Figure 2 and that the metrics for distance in this multi-dimensional space can be difficult to specify. For example, while some of the AMBR program participants seemed to consider the Challenge 3 problem to be fairly close to Challenge 2, we believe that the results (i.e., failure of all models to predict key results for Challenge 3) indicate that the Challenge 3 design point was actually quite distant from the Challenge 2 design point.

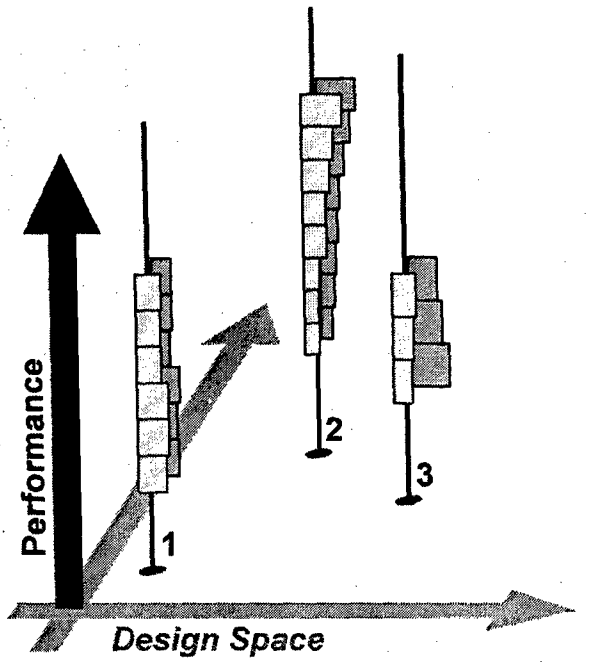

Figure 7. Validation Data Points in the Design Space (yellow = iGEN® AMBR model, green = human)

6.5 Developing and Validating CMs in the VE Context

The use of modeling and simulation in acquisition continues to expand – it supports earlier and cheaper evaluation of design alternatives in a broad range of contexts, from the smallest hardware component to the "system of systems." As noted above, the growing maturity of cognitive architectures and related tools seem to herald the use of operator models as a central element in the more organized approach to SBA recently promoted by various DoD directives and programs. However, our observations here suggest that cognitive models have not only failed to revolutionize the acquisition process but are currently a marginal consideration even when acquisition relies heavily on modeling and simulation and even when acquisition involves a highly operator-centric approach to system development. In part this is the result of major technology trends which, at least for the moment, have focused developers on other solutions to immediate problems of design evaluation.

The decades-long evolution of computer graphics technologies, resulting in both fast and cheap rendering of visual images, has both supported and driven the ways in which we now use and experience computer applications. Visualization has become pervasive, from graphical user interfaces in both desktop and Internet forms, to ever more realistic computer games. Of course it is no accident that military computer use is heavily invested in the same trend. DoD has supported basic research in computer graphics since the beginning and envisioned the use of virtual environments very early on, when a room full of computing power was required to allow a pilot to maneuver through a relatively crudely rendered three dimensional world. Many years of research and development in this area have resulted in the availability of relatively high fidelity virtual worlds at relatively low cost. This has allowed developers in the acquisition arena to push prototype testing into earlier stages of the design and development process through VE-

HITL – a "test pilot" can now fly a plane in various mission, threat, terrain, and weather contexts before a physical prototype is ever fabricated.

From the perspective of the acquisition manager, it is unlikely that any approach to design evaluation will instill more confidence in its results than one which employs human operators with experience related to the projected use of the system under development. Virtuality thus provides high confidence in design decisions, at low cost (when compared to building and testing actual prototypes). Reliance on virtuality in SBA seems to be limited only by the level of design detail required for modeling the system/component and the cost of the development/adaptation/integration efforts required to implement the system/component model and supporting elements in the virtual world (e.g., adversary platforms).

The trend to virtuality is by no means exhausted, currently being driven by various forces in both military and civilian contexts. In the military, in particular, concerns regarding training quality, cost, and timeliness are major forces pushing the technology forward. Although increased fidelity continues to be of interest (a prominent focus of commercial gaming), many major efforts in the military are related to infrastructure. In addition to allowing large groups of humans and simulations of various types to interact in a common context, the focus on large scale VE infrastructure also addresses issues of model and simulation reuse, a major cost factor for all uses of virtual environments, including SBA. It is fairly clear that over the next five to ten years these efforts will make the use of VE-HITL studies in SBA even more attractive based on increased fidelity, availability, and reduced costs. It is also clear that VE-based training will continue to expand, for similar reasons. Continued advances in virtuality will set the stage for the future development of CM use in SBA.

While increasing use of VE-HITL in SBA may continue to reduce the call for CMs to play the operator role in design evaluations (despite the need to address issues related to human variability and design sensitivity), widespread use of and improvements in virtual environment technologies also make CMs a more viable option for some purposes. The development of training systems which incorporate CMs to flesh out the required complement of friendly, neutral, and hostile roles in large scale VEs is clearly high on the list of likely near-term trends, especially based on the current pressures in the training arena. The development of reusable peripheral-role-playing agents will impact the acquisition process in several areas. It would seem a short step from the employment of cognitive models as peripheral role players to employing them to replace humans in primary operator roles for design testing.

Other than raising consciousness regarding the need for types of design testing which can only be accomplished realistically by means of simulation, moving forward on the use of cognitive models in SBA is primarily dependent on two factors: the cost associated with utilizing CMs and the level of confidence associated with results obtained from CM-based evaluations. Research efforts continue to address the cost of CM development through support for the continued development of existing cognitive architectures and tools, as well as for the development of new cognitive modeling approaches. These efforts will proceed independent of any specific SBA requirements, primarily motivated by needs in the training domain. Similarly, model reuse will continue to be a major focus of the development of both standards and VE infrastructure, largely independent of SBA. On the other hand, the level of confidence in CM-

based evaluation results for SBA is clearly tied to issues of validation, issues which may be unique to SBA.

Campbell & Bolton (Reference 58) have pointed out that outside of the use of cognitive models to advance cognitive science, application validity is the appropriate perspective for assessing cognitive models. In the case of simulation-based training, for example, cognitive models should not be determined to be valid on the basis of their ability to replicate human behavior, but rather on their ability to support effective training. If training through interactions with cognitive models improves human trainee performance, those cognitive models are valid for the training application. Campbell & Bolton note that, in the SBA context, the success or failure of the acquisition process itself might be considered the true test of application validity although this is of course not practicable. Instead, they suggest that CM validity may need to be conceived in more general terms, similar to construct validity. In any case, it is relatively clear that the "face validation" generally deemed adequate for the development of training applications is not likely to provide sufficient confidence in model results for major design decisions in the acquisition process (Reference 13). Although CM use in VE-based training will continue to expand and bring CM use into the acquisition process, developments in the training area are not likely to resolve the validation problem for SBA. Validation for SBA will need to occur in the context of SBA.

As described above, the validation of a cognitive model requires multiple data points in order to provide any confidence in its use for any other application than the one for which it was last calibrated. Figure 8 presents a concept for integrating multipoint CM validation within the acquisition cycle that is based on information collected on various ongoing large scale acquisition programs, as well as our assessment of the CM and VE trends just described. Moving from left to right, the acquisition timeline encompasses several stages including the pre-systems-acquisition, concept development stage, the systems acquisition proper, and the sustainment process which follows system deployment. The blue bar marks the period during which VE-HITL is utilized for concept development, design testing, and training, incorporating role player CMs, and possibly the same role player CMs throughout. As the system design evolves, the role player models need to evolve as well, perhaps assuming new tasks as the system is fleshed out. During this evolution, the role player CM will be calibrated on multiple occasions preferably based on data collected through the virtual environment on occasions when humans have played the role or roles in question. At the end of the acquisition process, it is also expected that the models will be calibrated based on data collected during operational T&E.

In general, calibration proceeds as a series of independent calibration events, each event requiring little reference to previous events. Calibration is accomplished with reference to the unique, current configuration of all components involved – the many components of the virtual environment as well as the cognitive model. In Figure 8, the green bar marks the period of expected analytic use of cognitive models in design evaluation through constructive simulation. The cognitive models involved may be identical to those used as role player agents, but here the issue of validity is addressed. At each validation point, all available human and simulation data collected from HITL and T&E to date, as well as any available data collected outside the acquisition program, is utilized to assess the validity of the model(s) over the design space. Based on this assessment, the validity of the model(s) for the current analytic application of

interest is determined. Obviously, confidence in analytic results will increase over time, particularly if the data points are distributed throughout the design space relative to the current design evaluation target. The availability of externally generated data points, particularly through calibration efforts which predate the acquisition process, may result in higher model confidence levels earlier in the acquisition process.

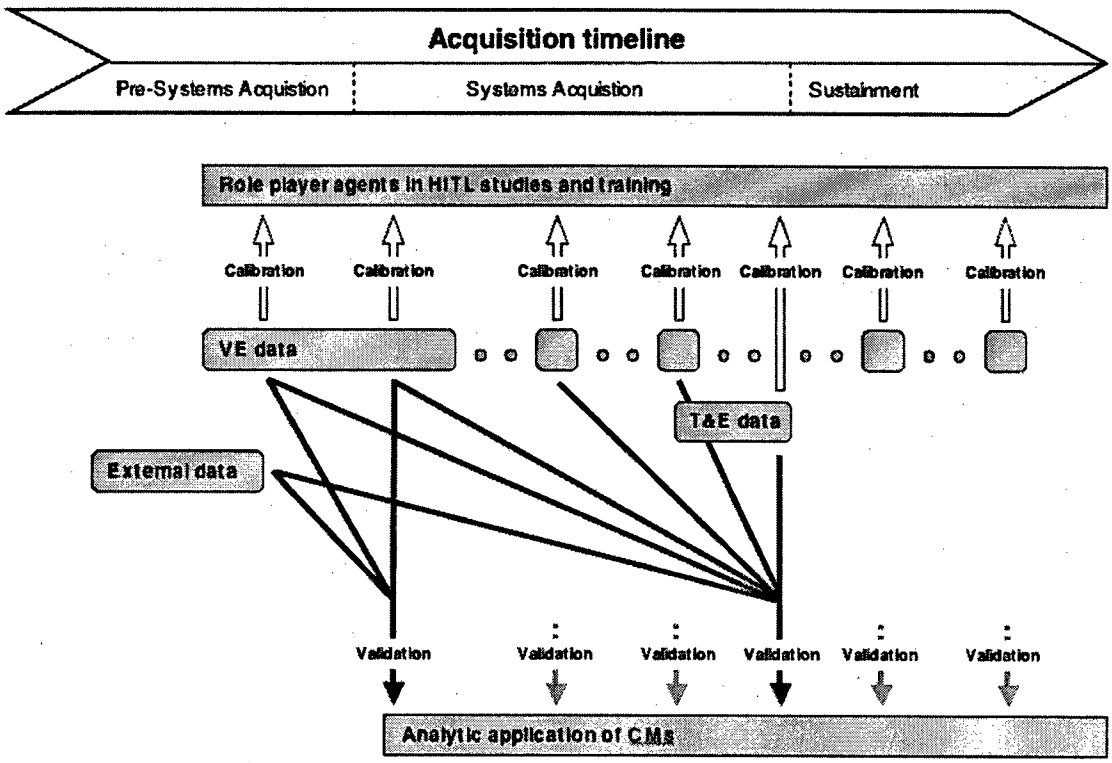

Figure 8. CM Validation in the SBA Process

7. Conclusions

Validation imposes both cost and scheduling constraints on the SBA process. Validation requires the involvement of both human subjects to generate comparison data and analysts to design the data collection process and analyze results. Since model validation must generally precede model use, an SBA manager could be forced to plan and execute a complex process involving the development of a test environment for data collection as well as the data collection itself, all in the early stages of the SBA process. Based on ongoing developments in SBA, we believe it is possible to integrate CM validation (and therefore CM use) into the larger SBA process, and ameliorate problems of cost and scheduling.

Although not being used directly to evaluate system designs, CMs are beginning to become visible in the acquisition process. Specifically, CMs are being used as peripheral role players (adversaries or teammates) in virtual environments (e.g., Reference 80). This application of CMs is emergent in three areas related to SBA:

- large scale training exercises which are often used in the concept development stage of SBA;
- simulation-based training under development for systems being acquired; and
- VE-HITL studies which are becoming pervasive in the SBA process.

As the use of CMs in these areas develops, the startup costs for CM-based design evaluation will decline. Not only may models similar, if not identical, to system operator models become available from these sources, but the environment and data sources for validation will also become available, particularly through VE-HITL. Working within the SBA process, use of and confidence in CMs can be bootstrapped by leveraging the results of the expanding role of virtual environments in SBA and the associated development of peripheral role player CMs.

Although we have developed a rather complex view of the appropriate considerations and processes for validation of CMs in the SBA context, we have also developed an expectation that appropriate assessments of the validity of CMs is achievable through empirically grounded analysis, with minimal reliance on art and intuition. This will require clear specification of the design space in which SBA decisions/assessments are to be made and the determination or development of validation data points that compare (pre-calibration) model to empirical performance data. The higher-level model specifications that we characterize as the CM architecture develop validation evidence from the collection of individual context-specific validations that are achieved with that architecture. And within each design-space context, the validity of the model at a new, untested design point is just a measure of the expected error that is estimated according to the most appropriate extrapolation from the prior validation data points to the new point of interest. Validation for synthetic agent applications in support of VE-HITL experiments is relatively easy; just a single calibration of the model to the desired design point will typically suffice as long as the design does not change across experimental cases from the perspective of the peripheral role of the synthetic agent.

8. References

1. Pew, R. W. & Mavor, A. S. (Eds). (1998). *Modeling Human and Organizational Behavior: Application to Military Simulations*, Washington, DC: National Academy Press.
2. Ritter, F.E., Shadbolt, N.R., Elliman, D., Young, R.M., Gobet, F., & Baxter, G.D. (2002). Techniques for Modeling Human Performance in Synthetic Environments: A Supplementary Review. Technical Report Number HSIAC SOAR 02-02. Wright-Patterson Air Force Base, OH: Human Systems Information Analysis Center.
3. Williams, M. (2001). Distributed Testing in Air Force OT&E. Presentation at the Georgia Tech Test & Evaluation Research and Education Center (TEREC) Testing for Information Assurance Conference, Albuquerque, NM. http://www.terec.gatech.edu/iaconf.html
4. VonHolle, J. (2004). The new SNA–revisited. *The MSIAC's Journal On-line*, (5)2. http://www.msiac.dmso.mil/journal/joe53_1.html [accessed July 15, 2004]
5. Siegel, A. I. & Wolf, J. J. (1969). *Man-machine simulation models*. New York: John Wiley.
6. Lane, N., M. Strieb, F. Glenn, and R. Wherry (1981). The human operator simulator: an overview. In *Manned Systems Design: Methods, Equipment, and Applications.* J. Moraal and K.-F. Kraiss (Eds). New York: Plenum Press.
7. Chubb, G.P. (1981). SAINT, A digital simulation language for the study of manned systems. In *Manned Systems Design: Methods, Equipment, and Applications*. J. Moraal and K.-F. Kraiss (Eds). New York: Plenum Press.
8. McRuer, D. and E. Krendel (1974). Mathematical models of human pilot behavior. AGARDograph 188. London, UK: NATO Advisory Group for Aerospace Research and Development.
9. Neville, K., N. Takamoto, J. French, S.R. Hursh and S. G. Schiflett (2000). The sleepiness-induced lapsing and cognitive slowing (SILCS) model: Predicting fatigue effects on warfighter performance. In *Proceedings of the 44th Annual Meeting of the Human Factors and Ergonomics Society* (pp 3-57 - 3-60). Santa Monica, CA: Human Factors and Ergonomics Society.
10. Allender, L., Kelly, T. D., Salvi, L., Lockett, J., Headley, D. B., Promisel, D., Mitchell, D., Richer, C. & Feng, T. (1995). Verification, validation, and accreditation of a soldier-system modeling tool. In *Proceedings of the Human Factors and Ergonomics Society 29th Annual Meeting* (pp. 1219-1223). San Diego, CA.
11. Brett, B.E., J.A. Doyal and C.R. Hale (2003). An analysis system relating individual human performance measures to overall mission effectiveness. In *Proceedings of the Twelfth International Symposium on Aviation Psychology*, Wright State University, Dayton OH. Clearance number ASC: 03-0347.
12. Gluck, K. and R. Pew. 2001. Lessons learned and future directions for the AMBR model comparison project. In *Proceedings of the Tenth Conference on Computer Generated Forces and Behavioral Representation* (pp.113-121). Orlando, FL: SISO, Inc.
13. Young, M.J. 2003. Human performance model validation: one size does not fit all. In *Proceedings of the Summer Simulation Conference.* San Diego, CA: Society for Modeling and Computer Simulation.

14. Harris, R., J. Bennett and J. Stokes (1982). Validating CAR: a comparison study of experimentally-derived and computer-generated reach envelopes. In *Proceedings of the Human Factors Society 26th Annual Meeting*, Santa Monica, CA: Human Factors and Ergonomics Society.
15. Moraal, J. & Kraiss, K.-F. (Eds). (1981). *Manned Systems Design: Methods, Equipment, and Applications*, New York: Plenum Press.
16. McMillan, G.R, Beevis, D., Salas, E., Strub, M.H., Sutton, R. & Van Breda, L. (1989). *Applications of human performance models to system design.* New York: Plenum Press.
17. Allender, L. (2000). Tools for modeling performance in systems through green-colored glasses: an Army perspective. In *Proceedings of the IEA 2000/HFES 2000 Congress,* San Diego, pp. 1-717 – 1-720.
18. Glenn, F., Barba, C., Ryder, J., Purcell, J., Weiland, M. & Convery, B. (1991). Review of Crew Station Development Tools for the Advanced Technology Crew Station Program. TR 911107.9000D10. Ft. Washington, PA: CHI Systems Inc.
19. Zachary, W., Campbell, G., Laughery, R., Glenn, F. & Cannon-Bowers, J. (2001). The application of human modeling technology to the design, evaluation, and operation of complex systems. In E. Salas (Ed), *Advances in Human Performance and Cognitive Engineering Research* (pp. 201-250). Amsterdam: Elsevier Science.
20. Atwood, M. E., Gray, W. D. & John, B. E. (1996). Project Ernestine: Analytic and empirical methods applied to a real-world CHI problem. In M. Rudisill, C. Lewis, P. B. Polson, and T. D. McKay (Eds). *Human-Computer Interface Design: Success Stories, Emerging Methods and Real-World Context* (pp. 101-121). San Francisco: Morgan Kaufmann.
21. Zachary, W., Le Mentec, J-C., Iordanov, V. (2001). Generating subjective workload self-assessment from a cognitive model. In *Proceedings of the Fourth International Conference on Cognitive Modeling.* Fairfax, VA: Erlbaum.
22. JSF (2004). http://www.jsf.mil/
23. Air Force (2004). http://www.airforce-technology.com/projects/jsf/
24. Alexander, D. (2004). Gearing up for F-35 assembly, *Aerospace Engineering*, April 2004. http://www.sae.org/aeromag/compengineering/04-2004/2-24-3-31.pdf
25. Machine Design (2003). Lockheed Martin Aeronautics Expands Simulation as It Plans Joint Strike Fighter Production in Texas, *Machine Design*, June 4, 2003. http://www.machinedesign.com/ASP/strArticleID/55861/strSite/MDSite/viewSelectedArticle.asp
26. Kwak, S.D., Andrew, E., Murtha, J., Brown, D. (2003). *Joint synthetic battlespace Desert Pivot Experiment (JPDE).* MITRE Technical Report, The MITRE Corporation.
27. Bowen, C.D., Couture, R.G., Flournoy, R.D., Forbell, E.M. & Means, C.M. (Aug 2002). *Capturing behavioral influences in synthetic C2: What we've learned so far and where we need to go.* MITRE Technical Report, The MITRE Corporation.
28. Leedom, D.K. (Oct 2004). *Project Gnosis: The Modeling of Sensemaking and Knowledge Management within a Joint Force C2 Network.* Available Online: http://www.ebrinc.com/ebr/ProjectGnosisSummaryPage.pdf
29. Mui, R.C.Y., LaVine, N.D., Bagnall, T., Sargent, R.A., Goodin, J.R. & Ramos, R. (2003). A method for incorporating cultural effects into a synthetic battlespace. In *Proceedings of the 2003 Conference on Behavior Representation in Modeling and Simulation.* Scottsdale, AZ.

30. Buker, G.E. & Flis, B.C. (2004). Use of intelligent controller nodes to augment human role-players in synthetic battlespace exercises. In *Proceedings of the 2004 Winter Simulation Conference* (pp 898-902). Washington, D.C.
31. Flournoy, R.D. (April 2002). *Leveraging Human Behavior Modeling Technologies to Strengthen Simulation-Based C2 Acquisition.* The MITRE Corporation.
32. Deutsch, S. & Cramer, N. (1998). OMAR human performance modeling in a decision support environment. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting.* Santa Monica, CA: HFES.
33. Stacy, W. (2004). *Application and Evaluation of Intelligent Agents for Performance Measurement Phase II* (Technical Report No. AP-R-1253). Woburn, MA: Aptima, Inc.
34. Nielson, P., Beard, J., Kiessel, J. and Beisaw, J. (2002). Robustness in behavior modeling overview. In *Proceedings of the 11th Conference on Computer Generated Forces and Behavioral Representation*, SISO, Orlando, FL: Institute for Simulation and Training.
35. Zachary, W. W. and Le Mentec, J.-C. (2000). Incorporating Metacognitive Capabilities in Synthetic Cognition Systems. In *Proceedings of the Ninth Conference on Computer Generated Forces and Behavioral Representation* (pp. 513-521). Orlando, FL: Institute for Simulation and Training.
36. Silverman, B.G., Cornwell, J.B. & O'Brien, K. (2003). *Progress to Date on the Human Performance Moderator Function Server (PMFserv) for Rapidly Generating Reusable Agents, Forces, and Crowds.* Available online: http://www.seas.upenn.edu/~barryg/HBMR.html
37. Glenn, F., Le Mentec, J.-C., Ryder, J., Santarelli, T., Stokes, J. and Zachary, W. (2003). Development of a Concept Learning Capability for a Human Performance Model. In *Proceedings of the 2003 Conference on Behavior Representation in Modeling and Simulation.* Scottsdale, AZ.
38. Van Lent, M., McAlinden, R., Brobst, P., Silverman, B.G., O'Brien, K. & Cornwell, J. (2004). Enhancing the behavioral fidelity of synthetic entities with human behavior models. In *Proceedings of the 2004 Conference on Behavior Representation in Modeling and Simulation.* Crystal City, VA.
39. Warwick, W. (2004). Building a Cognitive Menu within a CGF. *Paper presented at the Second Annual Workshop on Cognitive Systems.* Santa Fe, NM: Sandia National Laboratories.
40. Eggleston, R.G. (2004). Tightening the linkage of CSE and software systems engineering. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting.* Santa Monica, CA: HFES.
41. Sisti, A.F., Trevesani, D.A. & Reaper, J.H. (2003). Joint Synthetic Battlespace for Research and Development (JSB-RD). In A. Sisti & D.A. Trevesani (Eds), *Enabling Technologies for Simulation Science VII* (SPIE Proceedings Vol. 5091), Bellingham, WA: SPIE.
42. Trott, K. (2003). Command, Control, Communications, Computer, Intelligence, Surveillance, and Reconnaissance (C4ISR) Modeling and Simulation using Joint Semi-Automated Forces (Technical Report No. AFRL-IF-RS-TR-2003-144). Rome, NY: Air Force Research Laboratory.
43. Eirich, P. L., Coolahan, J. E. and Purdy, E. (2002). A Collaborative Environment Architecture for Future Combat Systems (FCS) Modeling and Simulation. *Spring*

Simulation Interoperability Workshop. (available at www.msiac.dmso.mil/sba_documents/Collab%20Environment%20for%20FCS%20M&S.pdf)

44. AMSO (2004). Army Modeling and Simulation Office (AMSO). http://www.amso.army.mil/
45. AMSO (2002). Planning Guidelines for Simulation and Modeling for Acquisition, Requirements and Training. September 20, 2002. http://www.amso.army.mil/smart/pol-guid/guidance/guidelines/guidelines-revisedsep02.doc
46. SMART (2004). http://www.peostri.army.mil/SMART/
47. SMART (2003). SMART Lessons Learned Case Summary: Aerial Common Sensor, July 23, 2003. http://www.amso.army.mil/smart/exec/lessons/ACS%20SMART%20LL%20Summary%20Jul%2003.doc
48. Army (1997). Army Regulation 5-11: Management of Army Models and Simulations Headquarters, Department of the Army, Washington, DC, 10 July 1997. https://134.11.61.26/CD4/Publications/DA/AR/AR%205-11%2019970710.pdf
49. Army (2004a). Warfighter Information Network - Tactical Simulation Support Plan Version 3.0, 28 May 2004, Department of the Army, Office of the Project Manager Warfighter Information Network – Tactical, Fort Monmouth, NJ. http://www.amso.army.mil/smart/exec/ssp/win-t/
50. Army (2004b). Simulation Support Plan: Modernized Longbow Apache Block III, ACAT level and Milestone Status Pending, 23 July 2004, Project Manager Apache, Redstone Arsenal, AL. http://www.amso.army.mil/smart/exec/ssp/longbow/
51. Stytz, M.R. and S.B. Banks (2003a). Progress and prospects for the development of computer-generated actors for military simulation: part 1—introduction and background. *Presence,* 12(3): 311–325.
52. Stytz, M.R. and S.B. Banks (2003b). Progress and prospects for the development of computer-generated actors for military simulation: part 3—the road ahead. *Presence,* 12(6): 629-643.
53. Banks, S.B. and M. Stytz. (2003). Progress and prospects for the development of computer-generated actors for military simulation: part 2— reasoning system architectures and human behavior modeling. *Presence*, 12(4): 422-436.
54. Zimmerman, W., Butler, R., Gray, V., Rosenberg, L. & Risser, D. T. (1984). Evaluation of the HARDMAN comparability methodology for manpower, personnel and training. Report No: NAS 1.26:173733; JPL-PUBL-84-10. Pasadena, CA: Jet Propulsion Laboratory.
55. Warner, N.W., Forster, E., Messick, M. & Wolf, J.J. (1995). Performance Metrics Methodology: Bridging the Gap between Subsystem Measures of Performance (MOPs) and Mission Measures of Effectiveness (MOEs). Presented at *the Eighth International Symposium on Aviation Psychology.* Columbus, OH.
56. Kieras, D.E. (2003). Model-based evaluation. In J.A. Jacko & A. Sears (Eds), *The Human-Computer Interaction Handbook.* Mahwah, NJ: Erlbaum. pp. 1139-1151.
57. Pew, R.W. & Gluck, K. (Eds) (2005). *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation.* Mahwah, NJ: Erlbaum.

58. Campbell, G. E. & Bolton, A. E. (2005). HBR validation: Integrating lessons learned from multiple academic disciplines, applied communities and the AMBR project. In *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation.* R.W. Pew and K. Gluck (Eds). Mahwah, NJ: Erlbaum.
59. DoD Modeling and Simulation Office (DMSO). (2001). VV&A Recommended Practices Guide Glossary. Washington, DC: Defense Modeling and Simulation Office. Retrieved April 1, 2004 from http://vva.dmso.mil/Glossary/Glossary-pr.pdf
60. Glenn, F., Cohen, D., Wherry, R. & Carmody, M. (1993). Development and Validation of a Workload Assessment Technique for Cockpit Function Allocation. In K. Hendy (Ed), *Proceedings of the Workshop on Task Network Simulation for Human-Machine System Design*, Farnborough, United Kingdom, 21 June 1993.
61. Ianni, J. 1999. A specification for human action representation. In *Proceedings of SAE International Conference on Digital Human Modeling for Design and Engineering.* Warrendale, PA: Society of Automotive Engineers, (CD-ROM).
62. Badler, N., J. Allbeck, L. Zhao and M. Byun. 2002. Representing and parameterizing agent behaviors. In *Proceedings of Computer Animation,* (pp 133-143). Geneva, Switzerland: IEEE Computer Society.
63. Fitts, P.M. & Peterson, J.R. (1964). Information capacity of discrete motor responses. *Journal of Experimental Psychology,* 67: 103-112.
64. Welford, A.T., Norris, A.H. & Shock, N.W. (1969). Speed and accuracy of movement and their changes with age. In W.G. Koster (Ed), *Attention and Performance II.* Amsterdam: North Holland.
65. Drury, C.G. (1975). Application of Fitts' Law to foot-pedal design. *Human Factors*, 17, 368-373.
66. Jagacinski, R.J., Repperger, D.W., Moran, M.S., Ward, S.L.& Glass, B. (1980) Fitts' law and the microstructure of rapid discrete movements. *Journal of Experimental Psychology: Human Perception and Performance*, 6, 309-320.
67. Brown, J.S. & Slater-Hammel, A.T. (1949). Discrete movements in the horizontal plane as a function of their length and direction. *Journal of Experimental Psychology*, 39, 84-95.
68. Topmiller, D. & Sharp, E. (1965). Effects of visual fixation and uncertainty on control panel layout (Technical Report AMRL-TR-65-149). Wright-Patterson AFB, OH: Aerospace Medical Research Laboratories.
69. Sargent, R.G. (1998). Verification and validation of simulation models. In *Proceedings of the 1998 Winter Simulation Conference* (pp 121-130). Washington, D.C.
70. Silverman, B.G., Might, R., Dubois, R., Shin, H., Johns, M. (2001). Toward a human behavior modeling anthology for developing synthetic agents. In *Proceedings of the 10th Conference on Computer Generated Forces and Behavioral Representation.* Orlando, FL: SISO.
71. Harmon, S. Y. (2002). A taxonomy of human behavior representation requirements. In *Proceedings of the 10th Conference on Computer Generated Forces and Behavioral Representation.* Orlando, FL: SISO.
72. Buff, W. L., Bolton, A. E. & Campbell, G. E. (2003). Providing an integrated team training capability using synthetic teammates. In *Proceedings of the American Society of Naval Engineers Human Systems Integration Symposium,* CD-ROM

73. Chapman, R. J., Ryder, J. and Bell B. (2004). STRATA (Synthetic Teammates for Real-time Anywhere Training and Assessment): An integration of cognitive models and virtual environments for scenario based training. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting*, New Orleans, LA. Santa Monica, CA: HFES.
74. Sargeant, R. (1999). Validation and verification of simulation models. In *Proceedings of the 1999 Winter Simulation Conference*, pp 39-48.
75. Deutsch, S., Diller, D., Benyo, B. and Feinerman, L. (2005). The Simulation Environment for the AMBR Experiments. In *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation.* R.W. Pew and K. Gluck (Eds). Mahwah, NJ: Erlbaum.
76. Gluck, K., Pew, R. and Young, M. (2005). Background, Structure, and Preview of the Model Comparison. In *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation.* R.W. Pew and K. Gluck (Eds). Mahwah, NJ: Erlbaum.
77. Diller, D., Gluck, K., Tenney, Y. and Godfrey, K. (2005). Comparison, Convergence, and Divergence in Models of Multi-tasking and Category Learning and in the Architectures Used to Create Them. In *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation.* R.W. Pew and K. Gluck (Eds). Mahwah, NJ: Erlbaum.
78. Pew, R., Gluck, K. and Deutsch, S. (2005). Accomplishments, Challenges, and Future Directions for Human Behavior Representation. In *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation.* R.W. Pew and K. Gluck (Eds). Mahwah, NJ: Erlbaum.
79. Zachary, W., Ryder, J., Stokes, J., Glenn, F., Le Mentec, J. & Santarelli, T. (2004). A COGNET/iGEN®Cognitive Model that Mimics Human Performance and Learning in a Simulated Work Environment. In *Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation.* R.W. Pew and K. Gluck (Eds). Mahwah, NJ: Erlbaum.
80. Scolaro, J. & Santarelli, T. (2002). Cognitive modeling teamwork, taskwork, and instructional behavior in synthetic teammates. In *Proceedings of the Eleventh Conference on Computer-Generated Forces and Behavior Representation*, (pp 315-322). Orlando, FL: UCF Institute for Simulation & Training.

9. Acronym List

ACE – Advanced Collaborative Environment

ACT-R – Atomic Components of Thought – Rational

AFRL – Air-Force Research Labs

AFRL/IF – AFRL Information Directorate

AMBR – Agent-based Modeling and Behavioral Representation

AMSAA – U.S. Army Materiel Systems Analysis Activity

AMSEC – Army Model and Simulation Executive Council

AMSO – Army Model and Simulation Office

AoA – analysis of alternatives

AOC – Air Operations Center

ASTARS – Army Standards Repository System

ATC – air traffic control

ATCS – Navy's Advanced Technology Crew Station

ATO – Air Tasking Order

AWACS – Airborne Warning And Control System

AWSIM – Air Warfare Simulation

BCC – Battle Control Center

C2 – Command and Control

C4ISR – Command, Control, Communications, Computer, Intelligence, Surveillance, and Reconnaissance

CAD – Computer Added Design

CADET – Computer Aided Design and Evaluation Techniques

CAFES – Computer Aided Function Allocation Evaluation System

CAR – Crewstation Assessment of Reach

CART – Combat Automation Requirements Testbed

CDDR – Concept Definition and Design Research

CE – Collaborative Environments

CGF – Computer Generated Forces

CM – cognitive model

CMC2 – Cultural Modeling for Command and Control

CMC2 – Cultural Modeling for Command and Control

CMMS – conceptual models of the mission space

COMBIMAN – Computerized Biomechanical Man-model

COP – Common Operating Picture

COTS – commercial off-the-shelf

CTA – cognitive task analysis

DARPA – Defense Advanced Research Projects Agency

DCOG – Distributed Cognition framework

DD(X) – multi-mission surface combatant ship program

DDS – decision support system

DIMSRR – Defense Intelligence Modeling and Simulation Resource Repository

DMO – Distributed Mission Operations

DMSO – Defense Modeling and Simulation Office

DoD – U.S. Department of Defense

DPD – Distributed Product Descriptions

DT&E – Developmental Test and Evaluation

EADSIM – Extended Air Defense Simulation

EASE – an ACT-R/Soar/EPIC hybrid

EDM – Engineering Development Models

EPIC – Executive Process Interactive Control

ETO – Effects Tasking Order

FCS – Army's Future Combat System

FCS – US Army's Future Combat Systems

FOM – Federation Object Model

FORTRAN – FORmula TRANslation

GOMS – Goals, Operators, Methods, and Selection rules

GOTS – government off-the-shelf

GSTF – Global Strike Task Force

GUI – graphical user interface

HARDMAN – Hardware vs. Manpower

HBM – Human Behavior Model

HBMA – Human Behavior Model Anthology

HBR – Human Behavioral Representation

HEDAD-O – Human Engineering Design Approach Document - Operator

HITL – human-in-the-loop

HLA – High Level Architecture

HOS – Human Operator Simulator

HPM – human performance model

HSI – human-system integration

HW – hardware

IADS – integrated air defense system

IMPRINT – Improved Performance Research Integration Tool

ISR – Intelligence, Surveillance, and Reconnaissance

JCIDS – Joint Capabilities Integration and Development Systems

JIMM – Joint Integrated Mission Model

JMEM – Joint Munitions Effectiveness Manuals

Joint STARS – Joint Surveillance Target Attack Radar System

JOSH – Joint STARS Operator Surrogate Human

JROC – DoD Joint Requirements Oversight Council

JSB – Joint Synthetic Battlespace

JSB-DS – JSB for Decision Support

JSB-RD – JSB for Research and Development

JSF – Joint Strike Fighter

JSIMS – Joint Simulation System

J-UCAS – Joint Unmanned Combat Air Systems

LCS – Littoral Combat Ship

LOE – level-of-effort

LSI – Lead Systems Integrator

M&S – Modeling and Simulation

MANPRINT – Army MANpower and PeRsonnel INTegration Program

MC2C – Multi-sensor Command and Control Constellation

MFS – U.S. Navy Manned Flight Simulator

MIDAS – Man-machine Integration Design & Analysis System

MODVAL – Human Performance MODel VALidation Program

MSRR – Army Model and Simulation Resource Repository

OT&E – Operational Test and Evaluation

PEO STRI – U.S. Army Program Executive Office for Simulation, Training, & Instrumentation

QFD – Quality Function Deployment

R&D – Research and Development

ROI – Return-On-Investment

RPICN – Role Player Intelligent Controller Node

SAINT – Systems Analysis of Integrated Networks of Tasks

SBA – simulation-based acquisition

SEAD – Suppression of Enemy Air Defenses

SMART – Simulation and Modeling for Acquisition, Requirements and Training

SME – subject matter expert

SOAR – State, Operator And Result

SRD – standards requirements document

SSE – Sum of the Squared Error

SSP – Simulation Support Plan

SUPPRESSOR – Suppressor Composite Mission

Simulation System

SW – software

T&E – Test and Evaluation

TCT – Time Critical Targeting

TTPs – tactics, techniques, and procedures

V&V – Validation and Verification

VE – virtual environment

VE-HITL – virtual environment human-in-the-loop

VV&A – Validation, Verification and Accreditation